MASTER'S THESIS

Handover and Channel Allocation Mechanisms in Mobile Satellite Networks

by Iordanis Koutsopoulos
Advisor: Leandros Tassiulas

CSHCN M.S. 99-10
(ISR M.S. 99-15)

ABSTRACT

Title of Thesis: HANDOVER AND CHANNEL ALLOCATION MECHANISMS IN MOBILE SATELLITE NETWORKS

Degree candidate: Iordanis Koutsopoulos

Degree and year: Master of Science, 2000

Thesis directed by: Professor Leandros Tassiulas
Department of Electrical Engineering

In this work we study first handover prediction in non-geostationary mobile satellite networks. The ultimate choice of the transition path depends on UT position and signal strength. We investigate the procedure of beam monitoring and propose UT maximum residence as the criterion for path selection.

The UT must operate both in full- and half- duplex mode, the latter being desirable when power limitations are imposed. We propose a scheme that achieves this goal and guarantees efficient diversity provision. Constant delay contours on the earth surface are defined. The problem of reliable time delay acquisition is addressed, in case synchronization is lost. The SBS solves that either by using

the known estimate of UT position or by requesting a measurement report by the UT.

The problem of channel allocation appears in cellular networks of every kind. Calls arising in the cell overlap area have access to channels of more than one base station and may choose which base station they will use to establish connection. In that case the problems of base station and channel assignment arise jointly.

We address the problem in a linear cellular network and aim at the minimum number of utilized channels. We present two algorithms: The first one expands Load Balancing in clique populations and is Sequential Clique Load Balancing (SCLB). The second one is named Clique Load Balancing with Inverse Water-Filling (CLB-IWF). In a dynamic environment, we unify SCLB and CLB-IWF into CLB-DA, which comprises Dynamic Allocation. CLB-DA is compared with Least Loaded Routing (LLR) policy and with Random Routing policy. We finally deduce that at light loads CLB-DA outperforms LLR, attaining smaller blocking probability, whereas at heavier loads all three policies converge.

HANDOVER AND CHANNEL ALLOCATION MECHANISMS IN MOBILE SATELLITE NETWORKS

by

Iordanis Koutsopoulos

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Master of Science
2000

Advisory Committee:

Professor Leandros Tassiulas, Chairman/Advisor
Professor Evaggelos Geraniotis
Professor Mark Shayman

Dedication

To my parents Georgios and Maria,

my sister Argiroula,

and to Asimina

for their everlasting and invaluable

love and support

Acknowledgements

I would like to express my honest thanks to my advisor, Professor Leandros Tassiulas for his guidance and invaluable comments throughout this work. I am also grateful to Dan Voce and Prabhat Kumar for their collaboration in the handover-related joint project with Hughes Network Systems. Their help in understanding practical system aspects and simulating respectively was great. Many thanks also to all people with whom I exchanged opinions and ideas on related matters and especially to my roommates Kyriakos Manousakis and Nikolaos Kanlis and also to Dr. Konstantinos Tsoukatos.

Table of Contents

List of Tables

List of Figures

HANDOVER AND CHANNEL ALLOCATION
MECHANISMS IN MOBILE SATELLITE
NETWORKS

Iordanis Koutsopoulos

February 11, 2000

This comment page is not part of the dissertation.

Typeset by LaTeX using the `dissertation` class by Pablo A. Straub, University of Maryland.

Chapter 1

Mobile Satellite systems for Personal Communications

1.1 Introduction

The tremendous growth of cellular telephone networks has demonstrated the demand for personal communications. The need for a broad range of telecommunication services (voice, data or image transmission) is growing and wireless access solutions are very appealing, since they provide users with mobility. The challenge lies in designing and implementing an efficient, low cost, personal communication service. Mobile satellite communication for commercial users is rapidly evolving towards Personal Communication Services (PCS) systems, capable of providing basic telephone, fax and data services essentially anywhere on the globe, at a per-minute cost which will be as little as three to four times as that of a regular wireless cellular service. Satellite systems offer the capability to provide location-insensitive, switched, broadband service, extending the reach of networks and applications anywhere on earth. To ensure seamless compatibility with those networks, a satellite system must be designed with the same essential characteristics as fiber networks, namely broadband channels, low error rates and low delays.

1.1.1 Geostationary and non-Geostationary Satellite systems

Satellite systems can be classified in two categories: Geostationary and Non-geostationary Earth Orbit satellite systems. A geostationary orbit is a circular orbit in the equatorial plane with an orbital period equal to that of the Earth, which is achieved at an orbital height of $35786km$. A satellite in a geostationary orbit will appear fixed above the surface of the earth. The footprint or service area of a geostationary satellite covers almost 1/3 of the earth's surface, so that near-global coverage can be achieved with a minimum of three satellites in orbit. At this height, communications through a GEO satellite entail a round-trip transmission delay of at least $250msec$. This GEO latency is the source of the annoying delay in many intercontinental phone calls. What can be an inconvenience on voice transmissions, however, can be implausible in real-time applications such as video-conferencing, as well as many standard data protocols-even the protocols underlying the Internet. Advanced digital broadband networks will be packet-switched networks, in which voice, video and data are all packets of digitized bits. It is not feasible to separate out applications that can tolerate delay from those that cannot. As a result, the network has to be designed for the most demanding application.

Non-geostationary satellite systems appear as an attractive solution to provide global coverage and ensure reliable communication for the most demanding applications. Two categories of orbits have been envisaged : Low Earth Orbit

(LEO) at an altitude of about 1,000 km and Medium Earth Orbit (MEO) at an altitude of about 10,000 km Subject to such orbits, the satellite moves continuously relative to the earth surface. Permanent global communications entail the use of several satellites, organized in constellations, with several satellites per orbit plane and several orbit planes per constellation. The traffic generated by a User Terminal (UT) is then supported by satellites successively passing over the service zone, and must be handed over from one satellite to the next.

1.1.2 Low Earth Orbit Satellite Systems

LEOs are either elliptical or (more usual) circular orbits at a height of less than $2,000km$ above the surface of the earth. The orbit period at these altitudes varies between 90 minutes and two hours. The radius of the footprint of a LEO communications satellite varies from 3000 to $4000km$. The maximum time during which a LEO satellite is above the local horizon for an observer on the earth is up to 20 minutes. A global LEO system requires a large number of satellites, in a number of different, inclined, orbits. When a satellite serving a particular user moves below the local horizon, it needs to be able to hand over the service to a succeeding one in the same or adjacent orbit. Due to the relatively large movement of a satellite in LEO with respect to an observer on the earth, satellite systems using this type of orbit need to be able to cope with large Doppler shifts. Satellites in LEO are also affected by atmospheric drag which causes the orbit to gradually deteriorate. However, LEO satellites offer the following advantages over other satellite systems:

- Global coverage provision.

- Multiple LEO satellite launches, due to small size of satellites.
- Reduced on-board and on-earth power requirements.
- Utilization of lightweight low power radio telephones with small low-profile antennas.
- Minimization of impact of time delay.

LEO satellite systems can alternatively be classified as follows:

- Little LEO : Orbcomm, VITA
- Big LEO : Iridium, Globalstar, ICO
- Broadband LEO : Teledesic

1.2 Big and Broadband LEO Mobile Satellite Systems

1.2.1 Iridium

From a technical standpoint, the Iridium LEO satellite system, proposed by Motorola and currently constructed by that company in conjuction with Lockheed Martin, Raytheon and other contractors is one of the most ambitious systems. The system is being purchased and operated by a separate company (Iridium, Inc.), which has secured investment from many parts of the world.

The design deploys 66 satellites, placed in circular polar orbits at $750km$ altitude. The satellites are deployed into six equispaced orbital planes, with 11 satellites equally separated around each plane. Satellites in adjacent planes are staggered with respect to each other to maximize their coverage at the equator, where a user may be required to access a satellite that is as low as 10^o above the horizon. LEO coverage is viewed as one offering low path delays and global coverage.

The system utilizes GSM-based telephony architecture and a geographically-controlled system access to the satellite. Eight users share $45msec$ transmit and $45msec$ receive frames, in channels that have a bandwidth of $31.5kHz$ and are spaced $41.67kHz$. In other words, users are synchronized, so that they transmit and receive in the same time windows alternately.

The Iridium system requires on-board processing to demodulate each arriving TDMA burst, and retransmit it to its next destination. If a Gateway Earth Station (GES) is in view, this can be accomplished on ground, otherwise it takes place on one of the following four nearest satellites: the one ahead or behind in the same orbital plane, or the nearest in either orbital plane to the east or west. The Inter-Satellite Links (ISLs) operate at 23 GHz, while links to the Gateway Earth stations are at 20 GHz. The utilization of ISLs greatly complicates system design, but allows global service provision with a small number of GESs and also gives more flexibility. A call can be routed within the satellite network and connected to any mobile located anywhere, or it can be connected to the public network through any GES. To properly route the traffic, each satellite must carry a set of stored routing tables from which new routing instructions are called every 2.5 minutes.

Provided services include voice and data at $2.4kb/sec$ and High Penetration paging, which affords $11dB$ more power than the regular signal. The design, however, already provides a link margin ($16dB$), which is higher than any of the competing systems. This is because Motorola required that the hand-held unit be usable from inside a vehicle, which in turn was dictated by the business plan, which depends heavily on serving international business travelers.

One of the complicating aspects of the Iridium system is the need to hand off a subscriber from beam to beam. Since a typical satellite pass takes less than 9 minutes there is also the need to hand off some calls to the next satellite to appear above the horizon, raising the possibility of the call being dropped, if buildings block the view. A further issue in Iridium system is the need to turn off beams as the satellites move away from the equator, to ensure that a subscriber can access only a single beam. Other non-geostationary systems attempt to exploit dual satellite visibility as a means of mitigating shadowing effects and claim that this is preferable to designing for high link margins [1].

1.2.2 Globalstar

The Globalstar system has been purchased by a limited partnership, in which Loral and Qualcomm are the principal partners. The satellites are built by Loral, while Qualcomm has developed much of the ground segment. Globalstar system will employ 48 satellites, organized in eight palnes of six satellites each. The constellation is designed for 100% single satellite coverage between $\pm 70^o$ latitude and 100% dual or higher satellite coverage between 25^o to 50^o latitudes. Globalstar will employ path diversity combining to mitigate blocking and shadowing. Each Globalstar satellite covers a comparable area of the earth's surface with only 16 spot beams.

The Globalstar system employs no ISLs and therefore a subscriber can gain access to the system while being visible by the satellite and a GES simultaneously. The system provides interconnection to Public Switched Telephone Network and Public Land Telephone Network (PSTN/PLMN) through 100 to 210 GESs for extension of terrestrial cellular call processing. Globalstar will sell access to the system to local service providers, who will have an exclusive regional right to offer service. Calls will only be established through satellite(s) when connections cannot be made over the terrestrial network.

Globalstar chose Qualcomm's terrestrial CDMA technology for the mobile link in order to increase capacity through frequency reuse and voice activity detection and to have the ability for spectrum sharing and improved multipath performance. Globalstar offers data rates at 1.2, 2.4, 4.8 and 9.6 kbps, and the vocoder rate is allowed to drop down to 1.2 kbps when no voice activity is detected. This reduces interference and increases capacity, while maintaining synchronisation. Globalstar's antennas are shaped for elliptical beams aligned with the satellite velocity vector to increase the time a user stays within each beam.

Since all 16 beams of all the 48 satellites are always active, each satellite in view of a subscriber will pick up the subscriber's signal and retransmit it in its feeder link. Thus, by tracking the several satellites in view of a given GES, two channels can be reserved open for the subscriber. The channel providing the stronger signal can then be selected for connection to PSTN. This feature should mitigate blocking by buildings and provides an automated "soft" handover from satellite to satellite [1].

1.2.3 ICO

ICO-Global is a spin-off from Inmarsat, which owns 15% of the corporation. The rest is presently owned by Inmarsat Signatories and Hughes, who is the builder of the spacecraft.

ICO has chosen an Intermediate Circular Orbit for its system, at an altitude of $10355km$ (changed to $10390km$, in late 1998) with 10 satellites and two spares arranged in two inclined circular orbits. The inclination of the orbit is 45^o, which reduces coverage at high altitudes, but allows for the lowest number of satellites. To improve the link margins on the ICO satellites, a design with 163 spot beams per satellite was chosen.

ICO considers the constellation choice as one providing high elevation angles, accommodating satellite spatial diversity and demonstrating acceptable propagation delay. The MEO altitude also provides for slow-moving satellites as seen from the earth, leading to fewer and simpler handover arrangements than in a LEO system. ICO also claims that the system's technical risk is acceptable, as the ICO system will be based on more mature and tested technologies. The chosen MEO constellation also allows system growth by adding planes as capacity requirements increase. TDMA was chosen for multiple access as ICO argues that it permits power-efficient modulation schemes, promising the ability to support peak traffic capabilities by increasing and switching the capacity within a beam to cover real life traffic distributions. Six subscribers are multiplexed into channels of 25.2 kHz in width, at a bit rate of $36kb/sec$. In this respect, ICO follows more closely the approach being adopted by Iridium. A disadvantage of this access scheme is that a soft handoff (e.g from beam to beam) is not automatic, and it is more difficult to exploit dual satellite visiblity. One method would be to send a burst via an alternate satellite and by noting the strengths of the regular and alternate paths, the subscriber terminal could determine which satellite presently affords the best path to the GES and could adjust its own burst time and frequency to select that satellite.

The ICO satellites are being built by Hughes Space and Communications Division and the ground segment by a team consisting of NEC, Ericcson, and Hughes Network Systems division. ICO hopes to have its system in operation in the 2000 to 2001 frame. The system plans to reuse as much as possible GSM technology in a narrowband TDMA satellite environment. User Terminals (UTs) are planned both as single mode and as dual mode, where the UT will work with both in the ICO standard and a regional terrestrial cellular standard (GSM in Europe, JDC in Japan, DAMPS in North America). ICO will be targeted primarily at users from the existing terrestrial cellular market, which travel to places where terrestrial cellular coverage is incomplete or non-existent. Road transport, maritime and aeronautical communities are also anticipated customers, in addition to the demand for semi-fixed applications in rural areas or in developing countries.

1.2.4 Teledesic

Using a constellation of LEO satellites, Teledesic and its international partners are creating the world's first network to provide affordable, worldwide, fiber-like access to telecommunications services such as computer networking, broadband Internet access, high-quality voice and other digital data needs. The Teledesic system consists of 12 planes of 24 active satellites at $1350km$ altitude. Teledesic aims to provide high data rate (broadband) fixed and mobile services, continuous global coverage, fibre-like delay and bit error rates less than 10^{-10}. Thus, rather than targeting at voice and supporting low-bit rate data for fax and messaging as the Big LEOs do, Teledesic focuses on providing wireless broadband services with a fibre-like quality, focusing on data and supporting voice. The term Broadband LEO is therefore more suitable for describing it.

Teledesic terminals communicate directly with the satellite network and support a wide range of data rates. The terminals also interface with a wide range of standard network protocols, including IP, ISDN, ATM and others. Although optimized for fixed-site terminals, the Teledesic network is able to serve transportable and mobile terminals, such as those for maritime and aviation applications. Most users will have two-way connections that provide up to 64 Mbps on the downlink and up to 2 Mbps on the uplink. Broadband terminals will offer 64 Mbps of two-way capacity.

Since the topology of a LEO-based network is dynamic, the network must continually adapt to these changing conditions to achieve the optimal, least-delay connections between terminals. The Teledesic Network uses a combination of destination-based packet addressing and a distributed, adaptive packet routing algorithm to achieve low delay and low delay variability across the network. Each packet carries the network address of the destination terminal, and each node independently selects the least-delay route to that destination. Packets of the same session may follow different paths through the network. The terminal at the destination buffers and if necessary reorders the received packets to eliminate the effect of timing variations.

Each satellite is a node in the fast-packet-switch network and has ISLs with other satellites in the same and adjacent orbital planes. This interconnection arrangement forms a robust non-hierarchical mesh which is tolerant to faults and local congestion. The network combines the advantages of a circuit-switched network (low delay "digital pipes") and a packet-switched network (efficient handling of multi-rate and bursty data). From a network viewpoint, a large constellation of interlinked switch nodes offers a number of advantages in terms of service quality, reliability and capacity. The richly interconnected mesh network is a robust, fault-tolerant design that automatically adapts to topology changes and to congested or faulty nodes and links. To achieve high system capacity and channel density, each satellite is able to concentrate a large amount of capacity in its relatively small coverage area.

The lowest frequency band with sufficient spectrum to meet Teledesic's broadband service, quality and capacity objectives is the Ka band. Downlinks operate between 18.8 GHz and 19.3 GHz, and uplinks between 28.6 GHz and 29.1 GHz. Communication links at these frequencies are degraded by rain and blocked by obstacles in the line-of-sight. To avoid obstacles and limit the portion of the path exposed to rain, the serving satellite must be at a high elevation angle above the horizon. The Teledesic constellation assures a minimum elevation angle of 40^o within its entire service area.

Within a cell, channel sharing is accomplished with a combination of Multi-Frequency Time Division Multiple Access (MF-TDMA) on the uplink and Asynchronous Time Division Multiplexing Access (ATDMA) on the downlink. Teledesic Network supports bandwidth-on-demand, allowing a user to request and release capacity as needed. This enables users to pay only for the capacity they actually use and for the network to support a much higher number of users [3]

Table 1.1: Comparative table of services and cost for some Mobile Satellite systems [2].

Service,cost	ICO	GLOBALSTAR	IRIDIUM	TELEDESIC
Service types	voice, data, fax, paging	voice,data,fax, paging, short msg. service, location	voice, data, fax, paging, location messaging	voice, data fax, paging, video
Voice(kbps)	4.8	2.4/4.8/9.6	2.4 / 4.8	16
Data (kbps)	2.4	7.2	2.4	16 - 2048
Modulation	QPSK	QPSK	QPSK	?

Voice ckts/sat	4500	2000 - 3000	1100(ltd)
			3840(max)
				0.1M 16 kbps channels

Dual-mode UTs	yes	yes	yes	no
Hand-held UTs	yes	yes	yes	portable
System cost (million $)	2600	2200	3700	9000
UT cost ($)	500	750	2500 - 3000	?
Satellite Lifetime	10 years	7.5 years	5 years	10 years
Call rates ($/minute)	1-2	0.35 - 0.53 wholesale	3	0.04 (16kbps min. service)
Operation scheduled	2000	1999	1998-1999	2003

Table 1.2: Comparative table of orbits and geometry for some Mobile Satellite systems [2].

Orbits, Geometry	ICO	GLOBALSTAR	IRIDIUM	TELEDESIC
Orbit class	MEO	LEO	LEO	LEO
Altitude (km)	10390	1410	780	695-705
Number of satellites	10 active 2 spares	48 active 8 spares	66 active 6 spares	840 active up to 84 spares
Number of planes	2	8	6	21
Inclination (deg.)	45	52	86.4	98.16
Period (minutes)	358.9	114	100.1	98.8
Sat. visibility time (minutes)	115.6	16.4	11.1	3.5
Min. UT elev. angle (deg.)	10	10	8.2	40
Min. mobile link prop. delay (msec)	34.5	4.63	2.60	2.32
Max. mobile link prop. delay (msec)	48.0	11.5	8.22	3.40
Min. GES elev. angle (deg.)	5	10	?	40
Number of Earth Stations	12	100-210	11	?
Coverage	global	within $\pm 70^o$ latitude	global	almost global (2^o hole at each pole)

Table 1.3: Comparative table of beams and reuse characteristics for some Mobile Satellite systems [2].

Beams and reuse	ICO	GLOBALSTAR	IRIDIUM	TELEDESIC
Access method	TDMA FDMA	CDMA FDMA	TDMA, F- & SDMA	(A)TDMA FDMA
Beams per satellite	163	16	48	64(supercells) 576 cells
Total num. of beams	1630	768	3168 (not all used)	53760
Beam diam.(km)	?	2254 avg.	600 (min.)	2.9 (cell)
Footprint diam. (km)	12900	5850	4700	1412
Satellite antenna	fixed,moving cells	fixed,moving cells	fixed,moving cells	earth-fixed cells
Reuse cells/clust)	4	1	12	9
Dual sat. visibility	≥ 2 satellites	"substantial"	at poles	mostly ≥ 2 satellites
sat. diversity ?	yes	yes	no	no (GES only)

Table 1.4: Comparative table of frequencies and miscellaneous characteristics for some Mobile Satellite systems [2].

Frequencies and miscellaneous	ICO	GLOBALSTAR	IRIDIUM	TELEDESIC
Mobile downlink freq. (MHz)	1980 - 2010	2483.5 - 2500 (S-band)	1616 -1626 (L-band)	Ka-band
Mobile uplink freq. (MHz)	2170- 2200	1610.0 - 1626.5 (L-band)	1616-1626 (L-band)	Ka-band
Feeder uplink freq.(GHz)	5 (C-band)	5.091 - 5.250 (C-band)	27.5 - 30 (Ka-band)	Ka-band
Feeder downlink freq. (GHz)	7 (C-band)	6.875 - 7.055 (C-band)	18.8 - 20.2 (Ka-band)	Ka-band
On-board processing ?	?	no	yes	yes
Inter-Satellite links (GHz)	N/A	N/A	22.5 -23.5	60
Handover ?	yes	yes, seamless	yes	yes
Link (fade) margin(dB)	8-12	11-16	16 voice 35 paging	?
Satellite output power (W)	2500	1000	1400	?
Satellite mass (kg)	1925	450	700	771

Chapter 2

Handover algorithms and modules for mobile satellite systems

2.1 Introduction

Existing terrestrial radio networks provide mobile communications services within limited regions. In order to extend the availability of these services and finally provide global coverage, several geostationary and non-geostationary satellite systems have been proposed as a supplement to these networks. In general, pure satellite systems can provide limited capacity in comparison to terrestrial networks, nevertheless they are particularly suited in order to cover large terrestrial areas with a scarce amount of traffic, since in these areas it is not convenient to implement cellular network equipment. Finally, satellite systems can be profitably used in order to cope with contingent situations of unavailability of terrestrial carriers (e.g due to traffic congestion in some cells) or in order to shorten the terrestrial tails (a mobile and a fixed user involved in a call, can communicate via satellite through the base station closest to the fixed user) [5] .

In order to provide communications services for small mobile or hand-held terminals with large enough elevation angle, the use of non-geostationary satellites is an appealing solution. Two categories of orbits have been envisaged [6]: Medium Earth Orbit (MEO) at an altitude of about 10,000 km, and Low Earth Orbit (LEO) at an altitude of about 1,000 km. Subject to such orbits, the satellite moves continuously relative to the earth surface, and permanent global communications entail the use of several satellites, organized in constellations, with several satellites per orbit plane and several orbit planes per constellation. The traffic generated by a User Terminal (UT) is then supported by satellites successively passing over the service zone, and must be handed over from one satellite to the next. Moreover, diversity attribute is provided as a means of mitigating unpredictable blockage during a call.

Several scenarios for efficient handover and resource allocation have prevailed in literature. In general, well established techniques and scenarios for terrestrial cellular systems can be expanded to satellite cellular systems. First of all, optimal scheduling of handovers is required to guarantee call quality of service. In [7] the handover problem is formulated as a stochastic optimization one, where the objective is to maximize an infinite horizon expected discounted reward obtained by the communicating mobiles minus a cost incurred for handovers. The reward is a function of some measurable characteristics of the received signal, such as signal strength, carrier to interference power ratios, channel fading, shadowing due to obstructions, propagation loss, power control strategies, traffic distributions, cell loading profiles, channel assignments, etc. Handovers are modelled as *switching penalties*, that are incurred because of the resources needed for their successful completion.

In the context of satellite handover, two strategies have been proposed [8]:

- Maximize the instantaneous elevation angle. In this strategy, always the satellite providing the higher elevation angle will be selected and handovers will be performed accordingly.
- Minimize the handover rate for a user, whereby the satellite that is visible with an elevation angle $\theta > \theta_{min}$ will always be chosen.

The standard procedure of beam signal level monitoring, applied in cell reselection schemes in GSM terrestrial cellular networks, is analyzed in [9] in the context of a mobile satellite system. The proposed system may be integrated or optimized in the presence of a positioning system (e.g GPS), however it can even work without that. In [10] a combined handover algorithm has been proposed, where transition decisions are dependent upon UT position and signal strength measurements.

One of the major problems in third generation mobile systems concerns the large amount of signalling information. In particular, because of the reduction of the beams' size and the presence of non-GEO satellites, the number of handovers tends to increase. In [11] it is mentioned that in satellite-fixed cell coverage systems the number of handovers experienced during a call is a function of the call duration, the beam size (beam handover), the satellite footprint size (satellite handover) and satellite speed, which depends on the orbit altitude. For the ICO satellite constellation, the satellite handover is typically one per hour, with possibly a beam handover every ten minutes or so.

In order to save the valuable satellite resources, signalling information must be kept to a minimum. In that respect, *seamless handover* is a smart approach for TDMA-based systems, since it does not interrupt the call and requires minimum signalling exchanges [12]. Seamless handover can be implemented by decentralizing several control functions at the Satellite Base Station (SBS). Each SBS, served by a certain satellite cell α initiates a handover procedure towards a new satellite cell b as soon as it perceives that the received power level relevant to cell α is below a certain threshold. In that case, the SBS detects the fraction of available timeslots within the frame associated to carriers assigned to cell b. If on one of those carriers f_c the fraction is larger than a given threshold (selected upon statistical considerations) it performs a handover towards cell b. Thus, it switches its transmitter to carrier f_c (while it still receives from the old cell α), until the network becomes aware of the handover and provides resources for routing the forward traffic via the new cell b.

Other practical approaches directly relate handover procedures and channel allocation techniques, where channel selection is performed according to the minimization of a channel assignment cost function [13]. Handover schemes can be broadly divided into prioritized or non-prioritized ones. In non-prioritized schemes, handover requests are treated in the same manner as originating calls and the probability of handover failure equals the probability of a call blocking. Handover prioritization schemes result in a decrease of the handover failure but they increase call blocking, which may reduce the total system capacity. This happens because channel assignment strategies allocate channels more readily to handover requests than they do to originating calls.

The Guard Channel concept offers a generic means of improving the probability of successful handover by reserving a fixed or dynamically adjustable number of channels exclusively for handovers [14]. The remaining number of channels are used for handovers as well as for originating calls. The penalty is the reduction of total carried traffic due to the fact that fewer channels are granted to originating calls. This disadvantage can be bypassed by allowing the queuing of originating calls. Intuitively, we can say that the latter method is feasible, because originating calls are considerably less sensitive to delay than handover requests. Another shortcoming of the employment of guard channels, especially with fixed channel assignment strategies, is the risk of inefficient spectrum utilization. Careful estimation of channel occupancy time distributions [15] is essential in order to minimize the risk by determining the optimum number of guard channels. With Dynamic Channel Assignment strategies, the SBS can reserve a set of channels only for handover requests, or it can have a number of channels with associated probabilities of being allocated for handover requests.

The Queuing of Handover requests has been analyzed in [16] and is based on the idea that when a mobile crosses a cell boundary in the direction of the neighboring cell, it will cross the overlapping area of both cells. While being in that area, the mobile can be served by either cell, which means that the handover can take place anywhere in this region. Consequently, the handover requests can be queued for a specific time period, equal to the time that the mobile traverses the overlap area. In the case of non-GEO satellite systems, this time depends on the adopted mobility model. When a call terminates, the system grants a channel to the call that has been waiting the longest, thus reducing forced termination probability.

A different perspective to efficient resource allocation is the implementation of dual satellite paths [17], [18]. A common two state channel model comprises *good state* and *bad state*. The former exists when there is no shadowing and can typically be characterized by a Rician distribution. The latter corresponds to a UT experiencing an obstruction of Line Of Sight (LOS) path to the satellite and can be characterized by Rayleigh/Lognormal distributed fades. Mitigation techniques such as increased power and bandwidth or modulation and coding are impractical means of obtaining the additional link margin required to ensure operation. A practical solution in a non-GEO satellite system is to establish and maintain a connection through two different satellites simultaneously. The satellites through which the connection is maintained, should at any time be located sufficiently apart from each other to minimize the probability that the LOS path to both satellites is obstructed at the same time.

We investigate the problem from a “macroscopic” point of view. In section 2.2 we build the basic setup and preliminaries that will be used throughout the simulation. In section 2.3 the basic algorithms for satellite and beam handover are analyzed and evaluated. Section 2.4 provides an insight into the sequence of events that occur prior to handover decision and turns attention into the maximum beam residence criterion for beam handover and section 2.5 focuses on the beam mobitoring procedure. In section 2.6 the diversity assignments taking place in the time of transition are presented. Finally in section 2.7 numerical results are illustrated and conclusions follow.

2.2 Preliminary structures and principles

2.2.1 System configuration

A simplified infrastructure of the satellite component of a mobile satellite system comprises the following parts:

1. n satellites $s_1, \ldots s_n$.

2. m beams per satellite footprint $b_1, \ldots b_m$.

3. r Satellite Base Stations (SBSs) $SBS_1, \ldots SBS_r$.

4. The earth globe, whose two-dimensional version is partitioned into squares with fixed longitude and latitude borders.

5. A traffic distribution percentage assignment p_{ij} to each of the above squares according to empirical data acquisition.

6. A population of mobiles (User Terminals, UTs).

For the case of the ICO mobile satellite system, the satellite constellation consists of $n = 10$ operational satellites moving at 10,355 km altitude and having a 6-hour revolution period. The satellites are arranged in two mutually orthogonal planes with 45^o inclination with respect to the equatorial plane, each plane comprising 5 operational satellites. The angular difference between two neighboring coplanar satellites is 72^o and the angle spacing between one satellite at its ascending node and an adjacent non-coplanar satellite at the beginning of simulation is 36^o. Each satellite of ICO provides mobile link coverage at $S-$band frequency through a set of $b = 163$ fixed spot beams with overlapping coverage. A spot beam has the shape of a cone emitted from the satellite and its projection on the earth forms the area served by the beam [19].

In ICO the $r = 12$ SBSs are connected to each other by a terrestrial network (inter-SBS network). In accordance to the terrestrial cellular model, each SBS consists of a Mobile Satellite Switching Centre (MSSC), a Land Earth Station (LES) and a Visitor Location Database (VLR) for mobility management. The LES on one side manages and relays the radio communications (via the satellites) to the UTs and on the other side relays communications to the MSSC for onward connection to the terrestrial networks. The MSSC provides a similar functionality with terrestrial cellular Mobile Switching Centres in managing and controlling the routing of calls between the UT and the terrestrial networks. Gateways (GW) outside the ICO infrastructure are connected by terrestrial links to the MSSC or a nearby SBS to provide the interconnection between the ICO infrastructure and the terrestrial networks. One or more Gateways are located in each country of the world.

2.2.2 Geographical coordinate systems

To record the position of a satellite, SBS or UT, the following coordinate systems are considered:

1. Orbital Plane system : This is the two-dimensional system of the satellite rotation plane, without considering plane inclination.

2. ECI (Earth Centered Inertial) System : This system is based at the center of the earth. The x-axis is fixed towards vernal equinox and the z-axis is the polar axis.

3. ECEF (Earth Centered Earth Fixed) system : This system is based at the center of the earth and rotates with it. The positive x-axis points always towards the intersection of the prime meridian and the equator (0^o longitude and 0^o latitude) and the z-axis is the polar axis.

4. OF (Orbit Frame): This satellite-based system is a newly introduced one. Its x-axis points in the direction of the satellite, the positive z-axis points towards the earth center and y-axis completes a right-handed triplet. This system provides a simple pictorial representation for the beam pattern,

overcoming complicated patterns on the curved earth surface and is the cornerstone of beam handover time prediction. A point on the earth surface is mapped onto a two dimensional system (the z-dimension gets eliminated), so that residence within a satellite footprint s_i and a specific beam b_j can easily be derived.

For completeness of arguments, the transformation method from the ECEF to the OF system are provided in the Appendix of this chapter.

2.2.3 Beam pattern

Getting as input the azimuth ϕ_i, elevation θ_i and half power beam width β_i angles relative to the OF coordinate system, each one of the 163 beams (cells)b_i is uniquely defined in the OF plane by a position vector of the beam center relative to the satellite nadir

$$\vec{W}_i = sin\theta_i cos\left(\phi_i + \frac{3\pi}{2}\right)\hat{x} + sin\theta_i sin\left(\phi_i + \frac{3\pi}{2}\right)\hat{y} \tag{2.1}$$

and its radius

$$R_i = tan\frac{\beta_i}{2} \tag{2.2}$$

2.2.4 Visibility concept

Let us denote by $\vec{S}(t)$ and $\vec{P}(t)$ the ECEF position vectors of the satellite and the UT respectively. A satellite is considered to be visible from the UT when its elevation angle relative to it,

$$\theta_{UT} = \cos^{-1}\left(\frac{-\vec{P}(t)\cdot\left[\vec{S}(t)-\vec{P}(t)\right]}{R_E\cdot|\vec{S}(t)-\vec{P}(t)|}\right) - \frac{\pi}{2} \tag{2.3}$$

is greater than 10^o. Similar visibility criteria hold for the SBS, the only difference being that the position vector of the SBS remains constant with time.

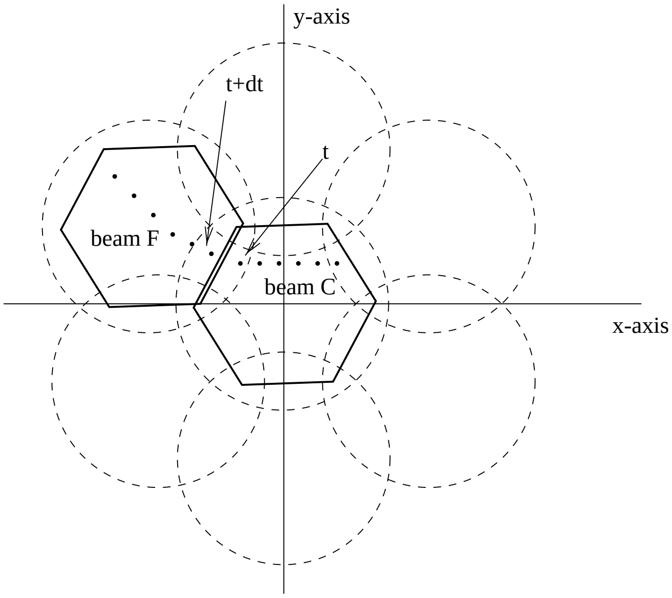

Figure 2.1: Demonstration of beam handover from the beam located in the satellite nadir to a neighboring beam.

2.3 Criteria for satellite and beam handover

The idea of satellite and beam handover time prediction is based on the principle of one-dimensional search within a finite time interval via bisection.

2.3.1 Beam Handover

The UT position is mapped to the OF using the transformation matrix ECEFtoOF, whose elements depend on current satellite position and velocity. Ephemeris data is used to determine future satellite locations, so that future positions of the UT in the OF are known. A binary search method of successive mappings of the UT position to the OF determines the time to handover to virtually any accuracy desired (other errors such as UT position, notwithstanding).

When a UT enters a beam, it is mapped to the OF several times using future satellite positions and velocities. Mappings are performed until a time interval of acceptable length (e.g. 1 minute) is found, where the UT resides in the current beam at the beginning of the interval and lies in a different beam at the end of the interval (Figure 2.1). Handover must occur sometime during this time interval and predicted handover time is selected as the midpoint of that interval. If the acceptable inaccuracy of the time interval is 1 minute, the handover prediction algorithm is accurate within 30 seconds. Equivalently, the contribution to the error in handover prediction by the prediction algorithm is at most 30 seconds in this case. Of course, in reality, several other factors contribute to prediction error, such as inaccuracy in ephemeris data and in UT position determination.

The speed of convergence of the algorithm depends on the initial horizon window length W_0 (in minutes) and the acceptable time prediction error, δ (in seconds). The number of required mappings in the OF coordinate system is then at most:

$$N = \left\lceil \log_2 \frac{60W_0}{\delta} \right\rceil \tag{2.4}$$

2.3.2 Satellite Handover

For the satellite handover we have the following procedure similarly:

- STEP 1: Obtain the serving satellite(s) at current time t_c.
- STEP 2: Update those to a future time t_f, using the satellite ephemeris data.
- STEP 3: If $\theta_{UT}(t_f) \leq 10^o$ or $\theta_{SAN}(t_f) \leq 5^o$, then conclude that a satellite handover has occurred at some time $t^* \epsilon \, (t_c, t_f)$.

- STEP 4: Apply the bisection idea on the appropriate interval.
- STEP 5: Stop after n iterations when $t_{f,n} - t_{c,n} < W$.
- STEP 6: The predicted satellite handover time is thus

$$t_s = \frac{t_{f,n} + t_{c,n}}{2} \tag{2.5}$$

2.4 Basic algorithms executed at the SBS

In order to ensure the most appropriate cell selection and a potential transition, each UT continuously monitors the received signal of the Broadcast Channel (BCCH) of a proper set of adjacent satellite cells and compares these qualities with that of the serving cell. This procedure allows the UT to perceive which adjacent cells are eligible to become serving cells and if the serving cell is still suitable. In case the serving cell is no more suitable, the UT triggers a cell reselection procedure, resulting in the selection of a new serving cell from this set of cells, according to some criteria.

2.4.1 Algorithm A: In-call BCCH selection for power measurements

The SBS periodically commands the UT to measure the BCCH signal strength of all serving and non-serving satellites in view of the UT. The SBS creates a list of all beams that will provide measurements back to the SBS later and serve as a confirmation to handover decisions. The list comprises a set $\mathcal{C}$ of beams, currently covering the UT position and belonging to visible satellites from both the UT and the SBS, and a set $\mathcal{A}$ of approaching beams of serving satellites.

The above two sets of beams are candidates for a satellite and a beam handover respectively. Upon reception of this list, the UT performs measurements for each of these beams and sends the enhanced list back to the SBS to be processed further. The procedure of finding the current beam is described as follows:

1. Translate the UT ECEF coordinates into the OF system, using the transformation matrix ECEFtoOF.

2. Set parameter *distance* to infinity.

3. Parse through all beams b_j. If the UT's distance from the beam center, r_j is less than the current *distance*, set current beam to b_j and update *distance* to r_j.

4. The output of the algorithm is the beam where the UT currently resides.

Taking into consideration the satellite footprint movement direction from right to left, a beam b_j is an upcoming beam for beam b_i if and only if it is located on the left of b_i, i.e $x_{j,OF} < x_{i,OF}$. A beam may have up to three approaching beams.

2.4.2 Algorithm B: Path Selection

The path selection algorithm provides input to resource allocation algorithm. The algorithm takes place after Algorithm A and before a handover of any type as well as a non-diversity to diversity transition attempt. Each entry of the list e_i is initially a pair of a satellite and a beam index (s_i, b_i). The list is modified as follows:

1. STEP 1: Take all possible combinations of single elements e_i of the list and append them at the end.

2. STEP 2: Eliminate the entries with a power measurement below a given threshold, indicating unreliable connection.

3. STEP 3: Eliminate double entries indicating a satellite in Diversity Reduction Mode, i.e in mode where diversity should not be donated.

4. STEP 4: Rank the list according to a predefined preference factor. At the end, the starting node of the list will have the highest preference factor.

Each entry of the list now represents either a single or a diversity path, eligible for resource allocation. The Preference Factor P_k is computed for every node k and is a function of the satellite elevation angle θ, the signal level I and the azimuth separation angle ϕ, in case the node denotes a diversity path. Thus

$$P_k = A \times \left(\frac{\theta_{k,1}}{\pi} + \frac{\theta_{k,2}}{\pi} + \frac{\phi_{k,12}}{2\pi} \right) + B \times (I_{k,1} + I_{k,2}) \tag{2.6}$$

In the above equation, I_k is an indicator, denoting whether the received signal level for the corresponding beam is above or below a given threshold level. In the former case $I_k = 1$, otherwise $I_k = 0$. The received signal strength is computed with a simplistic channel propagation model that takes into consideration the free space loss, the position of a UT in a beam, the multipath fading loss and the shadowing loss.

2.4.3 Maximum beam duration criterion for beam handover

From the analysis in the previous section, it became clear that the triggering factor for a handover was a combination of UT position and received signal power measurements. Motivated by the fact that additional handover occurances contribute to excess signalling load and significant transmission delays in the system, we propose a new criterion for handover event triggering. The basic characteristic is the minimization of the number of satellite and beam handover events, since the residence time in a cell is forced to be the maximum possible. Upon creation of the list with the candidate satellites and beams for transition, no preference factor computation is required. Simply the node containing the beam in which the mobile is predicted to stay the longest is selected as the transition beam. This beam may belong to the current serving satellite or not, providing thus the definition of beam handover or satellite handover after adopting this criterion.

This criterion requires less computation complexity than the corresponding path selection criterion. UT residence time for each beam in the list is computed by standard mappings in the OF satellite coordinate systemand no elevation angle computation is required. More important, no power measurement information exchange between the UT and the SBS is necessary in order to confirm handover decisions.

2.5 Estimation of BCCH measurement reception time

It was mentioned that the SBS commands the UT to measure the received signal strength of the BCCH bursts of a particular set of beams to support and confirm handover decisions. The list of candidate beams is created at the SBS and is transferred to the UT on an uplink Common Control Channel (CCH). Upon reception of the list, the UT performs measurements and returns the enhanced list to the UT on the downlink CCH. This procedure takes place both during signalling and traffic phase of a call.

In section 2.4.1 the issue of continuous measurement reports by the UT is discussed. The general idea is that the UT monitors the current beams of serving and visible non-serving satellites and the upcoming beams of serving satellites. A worst case scenario for this method would be a UT with diversity connection in view of four satellites, when the maximum number of measured beams would be 10.

The required time for the UT to perform measurements is proportional to the list size and the time point when this list is available back to the SBS is a function of this time. During signalling mode, it is highly recommended to minimize this time, so that transition to traffic mode is immediate. In traffic mode, this time is also critical for the SBS, in order to avoid outdated handover decisions, which may result in forced call termination. Therefore, an alternative method would be to command the UT to measure *only one* beam from every visible satellite, regardless if it is a serving one or not.

2.5.1 Assumptions

BCCH timeslot allocation and BCCH burst reception time

Broadcast Channels belong to the general category of common channels, which are shared between a number of users, in opposition to Traffic channels, which belong to Dedicated channels, carrying information for one user. Syncronization of common and dedicated channels is discussed in section 3.2.

A dynamic BCCH Frequency allocation plan is employed in realistic situations, where a BCCH frequency and timeslot is allocated to each beam and is constantly subject to changes. A relatively simplistic fixed allocation has been applied here, where the BCCH timeslot devoted to a specific beam is given by the formula

$$BCCH\_TN = b_i \ modN_b \tag{2.7}$$

where the beam index b_i takes values 1 to 163 and N_b is the number of BCCH timeslots in a BCCH frame.

A BCCH burst is considered to be eligible for measurement if and only if it is received by the UT for one timeslot duration in a *free* reference window.

Traffic timeslot allocation

We assume that the embedded TDMA scheme has $N_s = 6$ timeslots per frame, each of duration $T = 6.66msec$. In a real system, timeslots will be assigned to a UT according to resource allocation algorithms. Without loss of generality, we assume a random Traffic Channel (TCH) timeslot allocation, between 0 and $N_s - 1$. A UT which is assigned traffic timeslot x receives and transmits traffic during reference timeslots $x, x + 1, x + 2$, which form the reference window x.

Consider for simplicity that the UT is not provided diversity. Then, the windows where it receives and transmits traffic are

$$W_m = x + N_s \cdot m \ , m = 1, 2, \ldots \tag{2.8}$$

CCH message arrival at the UT

The list is tranferred to the UT via the CCH channel in one message, which is equivalent to 12 traffic bursts of $40msec$ each and its transmission time is considered to be practically independent of the list size.

2.5.2 CCH message arrival at the UT

Suppose that the list leaves the SBS at time $t = 0sec$. The time point when the UT receives the list depends on the relative positions of the SBS, the satellite and the UT. It can be approximated by

$$T_{r,UT} = d_{ss} + d_{c-s} + T_p + 12 \times 0.04 \ \ (sec) \tag{2.9}$$

where d_{ss} is the time delay from the SBS to the satellite, d_{c-s} is the C-band to S-band frequency conversion delay, and T_p is the propagation delay from the satellite to the UT.

Upon reception of the list, the UT intends to perform successive measurements of the BCCH bursts in the list. BCCH bursts leave the satellite in series at the beginning of a reference interval with a period of $N_b = 25$ timeslots. The beginning of the n-th BCCH burst arrives at the UT at times

$$t_n = [N_b(n-1) + BCCH\_TN] \times T + T_p \tag{2.10}$$

and is potentially subject to measurement by the UT, *only after* it has already accomplished the signal strength measurement of the previous burst.

Denote by $t_{n,k}$ the time instant when the beginning of the n-th series burst of the k-th element in the list is received by the UT, and by X_k the required time to measure the BCCH burst of the k-th element in the list. Then the above condition can be expressed as

$$t_{n,k} > T_{r,UT} + \sum_{\ell=0}^{k-1} X_\ell \tag{2.11}$$

2.5.3 BCCH measurement procedure and list return to the SBS

Assume that the beginning of the BCCH burst reception occurs at timeslot x_{begin}. Then the timeslot index when it ends, x_{end} is $x_{end} = x_{begin} + 1$, except when the arrival time coincides with a timeslot beginning, in which case, $x_{begin} = x_{end}$. Fix attention on the n^*-th BCCH series at time $t_{n^*,k}$. Then,

$$\begin{aligned} x_{begin,n^*} &= int\left(\frac{t_n}{T}\right) \\ &= N_b(n-1) + BCCH\_TN + int\left(\frac{T_p}{T}\right) \end{aligned} \tag{2.12}$$

BCCH bursts will arrive sequentially, until the existence of an open reference window (three empty consecutive slots) is detected. The r-th order BCCH of the n^* BCCH series arrives at the UT at times

$$A_{begin,n^*,r} = x_{begin,n^*} + N_b \times (r-1) \tag{2.13}$$

and ends at timeslot $A_{end,n^*,r} = A_{begin,n^*,r} + 1$ or at $A_{end,n^*,r} = A_{begin,n^*,r}$. To find an unoccupied reference window, the order of BCCH, r^* must be such that

$$\begin{aligned} A_{begin,r^*} &\neq x + N_s \cdot m \\ A_{begin,r^*} &\neq x + N_s \cdot m + 1 \end{aligned} \tag{2.14}$$

$$
\begin{aligned}
A_{begin,r^*} &\neq x + N_s \cdot m + 2 \\
A_{end,r^*} &\neq x + N_s \cdot m \\
A_{end,r^*} &\neq x + N_s \cdot m + 1 \\
A_{end,r^*} &\neq x + N_s \cdot m + 2
\end{aligned}
\tag{2.15}
$$

List returned to the SBS

Measurements are assumed to be performed instantaneously. The tranfer time of the list with the measurements back to the SBS is

$$T_2 = T_p + d_{s-c} + d_{ss} + 12 \times 0.04 \ (sec) \tag{2.16}$$

As a consequence, the *total time* needed for the SBS to recover the list of measurements, will be

$$T_{total} = T_{r,UT} + T_2 + \sum_{k \ in \ list} X_k \tag{2.17}$$

2.5.4 Proposed procedure

The method of measuring only one beam for each visible satellite has the obvious advantage of requiring less time to accomplish. The SBS receives the list and proceeds in allocating resources to the call, so that it enters traffic mode. Therefore, this method can be applied during signalling mode, where a fast switch to an initial path is required. However, the method may result in the selection of a suboptimal path if applied in traffic mode. The proposed method to estimate the time the SBS recovers the enhanced list, is summarized in the following algorithm:

1. Compute the time needed for the UT to receive the list, $T_{r,UT}$.

2. For every beam element of the list do

    - Find the time when the first BCCH arrives at the UT after completion of previous measurements by the UT (or after the arrival of the CCH message, for the first element in the list)
    - Find the absolute timeslot number (ATN) where this time corresponds and store it as a reference for the upcoming BCCHs.
    - Keep increasing the order of the arriving BCCH until an unoccupied timeslot is found.
    - Add this time to the measurement time for that beam

3. Compute the transmission time for the list to the UT.

4. Find the total elapsed time for the SBS to recover the list, as the sum of the times calculated from steps 1-3.

2.6 Diversity

Diversity is required in the ICO system to provide call continuity where there exists an unpredictable blockage of the mobile link to the UT. The diversity attribute includes the initial setting up of the dual path, the handover from of one or both paths, the reselection of the best of the two paths to be utilized for communications and the relinquishing of one path when necessary. Due to resource consumption, diversity may be donated to a certain percentage of active calls. Because satellite capacity is finite, a satellite may enter Diversity Reduction mode when the percentage of diversity paths through that satellite exceeds a configurable threshold.

2.6.1 Diversity assignments

Given that a satellite or beam handover has been predicted, the ranked list from Path Selection algorithm is recalled. The call is routed to the appropriate path and resource are allocated to the call. It is possible though unlikely, that the list is empty either at the beginning of a call (signalling mode) or in traffic mode, for the reason that there exists no visible satellite from both the UT and the SBS. In the former case the call is not initiated and in the latter case, it is forced to terminate. Since ICO satellite system guarantees global coverage at any time, this possibility has been excluded.

Consider a call which is currently either in diversity status or utilizes a single path and there exists resource margin, so that diversity can be donated to it. The most preferable transition target will then be the diversity node of the list having a high preference factor and a common path with the serving diversity node or single path. To this end, the list is examined, until a node satisfying the above matching condition is found. If we assume that the current serving satellite diversity pair is (A, B), the matching condition is satisfied if the list contains a pair of the form (A, C), (B, C), (C, A) or (C, B). The above reasoning is also applied in the case of a single call. The assignment of the new diversity entry is implementable in such a way, that any matching parts remain unaltered. The assignment method is extremely important for the derivation of correct handover rate, considering for example the transition $(A, B) \rightarrow (C, A)$ as one transition and not as two. More specifically:

1. If the current serving pair is $(Serv_1, Serv_2) = (A, B)$ and the target pair in the list is (A, C), then do the assignment $Serv_2 \rightarrow C$.

2. If the current serving pair is $(Serv_1, Serv_2) = (A, B)$ and the target pair in the list is (B, C), then do the assignment $Serv_1 \rightarrow C$.

3. If the current serving pair is $(Serv_1, Serv_2) = (A, B)$ and the target pair in the list is (C, A), then do the assignment $Serv_2 \rightarrow C$.

4. If the current serving pair is $(Serv_1, Serv_2) = (A, B)$ and the target pair in the list is (C, B), then do the assignment $Serv_1 \rightarrow C$.

We denote all these cases where a path exists as *diversity additions.* In the particular case no matching node is found, then the highest diversity entry is selected, but the diversity path (A, B) is switched to (C, D). This kind of diversity allocation is referred to as *new diversity.* Evidently, the majority of diversity assignments will be implemented under diversity addition. There also exists the possibility that a single path is most preferable than diversity paths, in which case the UT acquires a single path. When the percentage of calls has exceeded a certain threshold percentage, no diversity paths are eligible for transition. Non-diversity calls are periodically updated to determine if they have become eligible for diversity, considering the overall diversity percentage. Therefore, it is obvious that when the diversity monitoring time point t_m is very close to a handover event time t_h of a single path call, then this node containing the only path does not exhibit a high enough preference factor. Hence, another node with two different paths is very likely to qualify for a target path, resulting to a *new diversity* addition. The ultimate goal is maximum resource utilization either by diversity assignments or single path allocations, resulting at maximum throughput (maximum number of served calls).

2.7 Simulation and results

Aiming at a simulation which will generate realistic results, a real satellite system environment has been built. First of all, a representative traffic distribution has been adopted for the busy hour of the system, i.e the time when the traffic reaches its maximum. A total population of $S = 2.4M$ subscribers is assumed and a (configurable) percentage $q = 15\%$ states the number of UTs which are likely to become active during the busy hour. According to the given geographic distribution of call arrivals, specific terrestrial areas expose greater traffic density, whereas others (e.g. the poles, or areas covered by sea) are characterized by negligible traffic. The earth surface is projected onto a two dimensional plane and is divided into 288 $15^o \times 15^o$ squares, covering a surface from -180^o to 180^o longitude and 90^o to -90^o latitude. More specifically, the square with coordinates (i, j) is associated with the square area of

$$Longitude\ \epsilon\, [-180 + 15(j-1), -180 + 15j] \tag{2.18}$$

$$Latitude\ \epsilon\, [90 - 15(i-1), 90 - 15(i-2)] \tag{2.19}$$

for $i = 1, 2, \ldots 12$ and $j = 1, 2 \ldots 24$. Calls are assumed to arrive in independent Poisson streams with means

$$\lambda_{ij} = \frac{q \times S \times p_{ij}}{3600} \tag{2.20}$$

while call hold times follow the exponential distribution with mean $150sec$.

The simulation was coded in C++ and several numerical results serve as an estimation to system parameters. Results are obtained either on a local basis (i.e for a region of certain longitude and latitude range) or globally, for the entire earth surface. Consider a point $(\theta_{lat}, \theta_{elev})$ in the first graph of figure 2.2. This comprises all squares with coordinates (i, ℓ) with $\ell = 1, 2, \ldots 24$ and i fixed, or in other words is identical to the zone between $90 - 15(i-1)$ and $90 - 15(i-2)$ degrees latitude. If N_ℓ mobiles are generated at each such square during the busy hour and each of those mobiles k_ℓ with call duration τ is characterized by an average satellite elevation angle

$$\overline{\theta}_{k,\ell} = \frac{1}{\tau} \int_0^\tau \theta_{k,\ell}(t) dt \tag{2.21}$$

then, θ_{elev} is derived as follows:

$$\theta_{elev} = \sum_{m:N_m \neq 0} \overline{\theta}_{m,\ell}, \;\; m = 1, 2, \ldots 24 \tag{2.22}$$

Angles are averaged over those $15^o \times 15^o$ squares which contain non-zero traffic. From figures 2.2 and 2.3, it is perceived that the average satellite elevation and satellite azimuth separation angles obtain values within a certain range. Depending on the particular geographical location on earth, the elevation angle varies between 30^o and 48^o and the azimuth separation angle varies between 65^o and 135^o on the average. Satellites with high elevation angle and with large separation angle (in the case of diversity) are preferable for transition. In the simulation the contribution of those parameters and power measurements is taken into consideration, i.e $A = B = 1$ in equation 2.6.

Regarding diversity path allocations, the approximate percentages of *diversity additions* and *new diversity* paths are depicted in figure 2.4.The diversity monitoring time interval is 1 minute, while diversity threshold percentage is set to 40%. The results are expected, namely that diversity allocation is mostly realized as diversity path addition after Path Selection algorithm. Diversity additions concern at least the 88% of the assignments and, in some cases they can be considered as the only diversity assignment mode. Since a new diversity path allocation depends on the proximity of the diversity monitoring time point and the handover time instant of a single path call, by varying the diversity monitoring period, we obtained different percentages, lying within the same range $(88\% - 100\%)$.

In figures 2.5 and 2.6 we present some comparative results for the position and the maximum beam residence criterion about satellite and beam handover rates in two regions having moderate and heavy load (0.92 and 4.87 calls per second) respectively. By using maximum residence criterion, a reduction to beam handover rate up to $35 - 40\%$ has been observed in the steady state for a region with heavy traffic, while the reduction can be even greater (up to $85 - 90\%$) for mild traffic load. A small drawback in that former case may be the increased satellite handover rate by some amount for specific time periods. Taking into consideration the low satellite handover rate (around $3-4$ handovers per minute in steady state), this fact should not receive further concern. At any rate, under heavy traffic, the satellite handover rate is reduced by more than 50% as well.

In figure 2.7 the satellite coverage percentages as a function of latitude on the earth surface are depicted. It is confirmed that the ICO system provides global coverage, as for each location on earth there always exists a visible satellite. The possibility of two satellites being visible simultaneously is also considerably high (more than 80%), providing the system with the capability to establish diversity. For low latitude values, even three satellites are visible sometimes $(40\% - 70\%)$, and in general approximately one in three or four calls is covered by three satellites. For selected latitude regions $10^o - 30^o$, four satellites are also visible simultaneously.

Finally, results concerning the estimated list recover time by the SAN have been obtained by performing the experiment for two dfferent SBS locations of earth

- Usingen, Germany at $50^0 19' 52'' N$ latitude and $8^0 28' 25'' E$ longitude.
- Chattarpur, India at $28^0 31' 27'' N$ latitude and $77^0 11' 12'' E$ longitude.

To each of the squares $s_{ij} \equiv (i, j)$ in the geographic configuration, we associate an average time $\overline{T}_{ij}$, a minimum time m_{ij}, a maximum time M_{ij} and a time variation $Var_{t,ij}$. Over the entire earth, we define the average T, average minimum m,average maximum M and the average variance Var_t of measurement recovery time. We also define the absolute minimum and maximum times as

$$m_\alpha = \min_{i,j} m_{ij} = \min_{i,j} \min_{\beta \epsilon s_{ij}} T_{total,\beta} \tag{2.23}$$

$$M_\alpha = \max_{i,j} m_{ij} = \max_{i,j} \max_{\beta \epsilon s_{ij}} T_{total,\beta} \tag{2.24}$$

and β is the UT index.

The statistics per square are obtained by averaging over the total number of measurements performed by all UTs in that geographic location. The measurement results are summarized in tables 2.1 and 2.2. We notice that the involved average elapsed times are virtually independent of the location of the SBS. The measurement method using only one beam per satellite reduces the measurement recovery time by 60% on the average. Average minimum and maximum times are reduced accordingly, as well as the variance of the elapsed time. Taking into account the specifications about the maximum tolerance in waiting time during call set up, the percentage of cases when only one beam per satellite is used can be derived. All beams must be measured during traffic phase and the method of one beam per satellite should be used only exceptionally.

2.8 APPENDIX A

Transformation matrix between ECEF and OF coordinate systems

The transformation matrix will be a 3×3 matrix that relates the ECEF and OF coordinate systems and is obtained as follows

- STEP 1: Get the satellite position and velocity vectors in the ECEF coordinate system, $\vec{P}(t)$ and $\vec{V}(t)$ and the corresponding magnitudes $|\vec{P}(t)| = R_E + h$ and $|\vec{V}(t)| = 2\pi(R_E + h)/T$.

- STEP 2: Evaluate the vector $\vec{\gamma}(t) = -\vec{P}(t) \times \vec{V}(t)$ and the magnitude $|\vec{\gamma}(t)|$.

- STEP 3: Get the third row of the transformation matrix as:

$$\vec{r_3} = -\frac{\vec{P}(t)}{|\vec{P}(t)|} \tag{2.25}$$

- STEP 4: Get the second row of the transformation matrix as

$$\vec{r_2} - \frac{\vec{\gamma}(t)}{|\vec{\gamma}(t)|} \tag{2.26}$$

- STEP 5: Get the first row of the transformation matrix as

$$\vec{r_1} = \vec{r_2} \times \vec{r_3} \tag{2.27}$$

2.9 Appendix B

Computation formulas for the azimuth separation angles

The azimuth separation angle is the angle on the surface of the earth between arc L_1 which connects the UT position and the subsatellite point of the first satellite and arc L_2 which connects the UT position and the subsatellite point of the second satellite. The azimuth separation angle is inserted as a deciding parameter in the path selection algorithm in the case of a diversity path and has substantial importance in selecting the final path. For fixed elevation angle and power measurements, a big azimuth angle provides a preferable path, since there are fewer chances that both paths will be corrupted from an unpredictable blockage.

The function $D(\cdot,\cdot)$ computes the distance of two points with given longitudes and latitudes $P_1(\phi_1, \theta_1)$ and $P_2(\phi_2, \theta_2)$ on the earth globe as

$$D(P_1, P_2) = 2\tan^{-1}\left(\frac{\sin\alpha_x}{\sin\alpha_y}\ \tan\frac{a_1 - a_2}{2}\right) \tag{2.28}$$

where

$$\alpha_x = arctan\left(\frac{\cos((a_1 - a_2)/2)}{\cos((a_1 + a_2)/2)} \times \frac{1}{\tan(a_{12}/2)}\right) \tag{2.29}$$

$$\alpha_y = arctan\left(\frac{\sin((a_1 - a_2)/2)}{\sin((a_1 + a_2)/2)} \times \frac{1}{\tan(a_{12}/2)}\right) \tag{2.30}$$

and $a_1 = \pi/2 - \theta_1$, $a_2 = \pi/2 - \theta_2$,

$$a_{12} = \begin{cases} \phi_1 - \phi_2 & \text{if } \phi_1 - \phi_2 < \pi \\ 2\pi - (\phi_1 - \phi_2) & \text{otherwise} \end{cases} \tag{2.31}$$

The azimuth angle is then calculated using standard formulas from spherical trigonometry:

$$\phi_{azim} = \frac{\cos(D(S_1, S_2)) - \cos(D(S_1, UT))\cos(D(S_2, UT))}{\sin(D(S_1, UT))\sin(D(S_2, UT))} \tag{2.32}$$

Table 2.1: Statistical results over the entire earth for the mesurement recovery time using the two alternative methods, by an SBS located in Germany.

Quantity	Entire list	One beam per sat
T	3.405 sec	2.063 sec
m	1.741 sec	1.258 sec
M	5.506 sec	3.167 sec
Var_t	0.597 sec	0.159 sec
m_α	1.432 sec	1.135 sec
M_α	7.283 sec	4.221 sec

Table 2.2: Statistical results over the entire earth for the mesurement recovery time using the two alternative methods, by an SBS located in India.

Quantity	Entire list	One beam per sat
T	3.410 sec	2.063 sec
m	1.718 sec	1.255 sec
M	5.518 sec	3.188 sec
Var_t	0.583 sec	0.157 sec
m_α	1.380 sec	1.122 sec
M_α	7.442 sec	4.244 sec

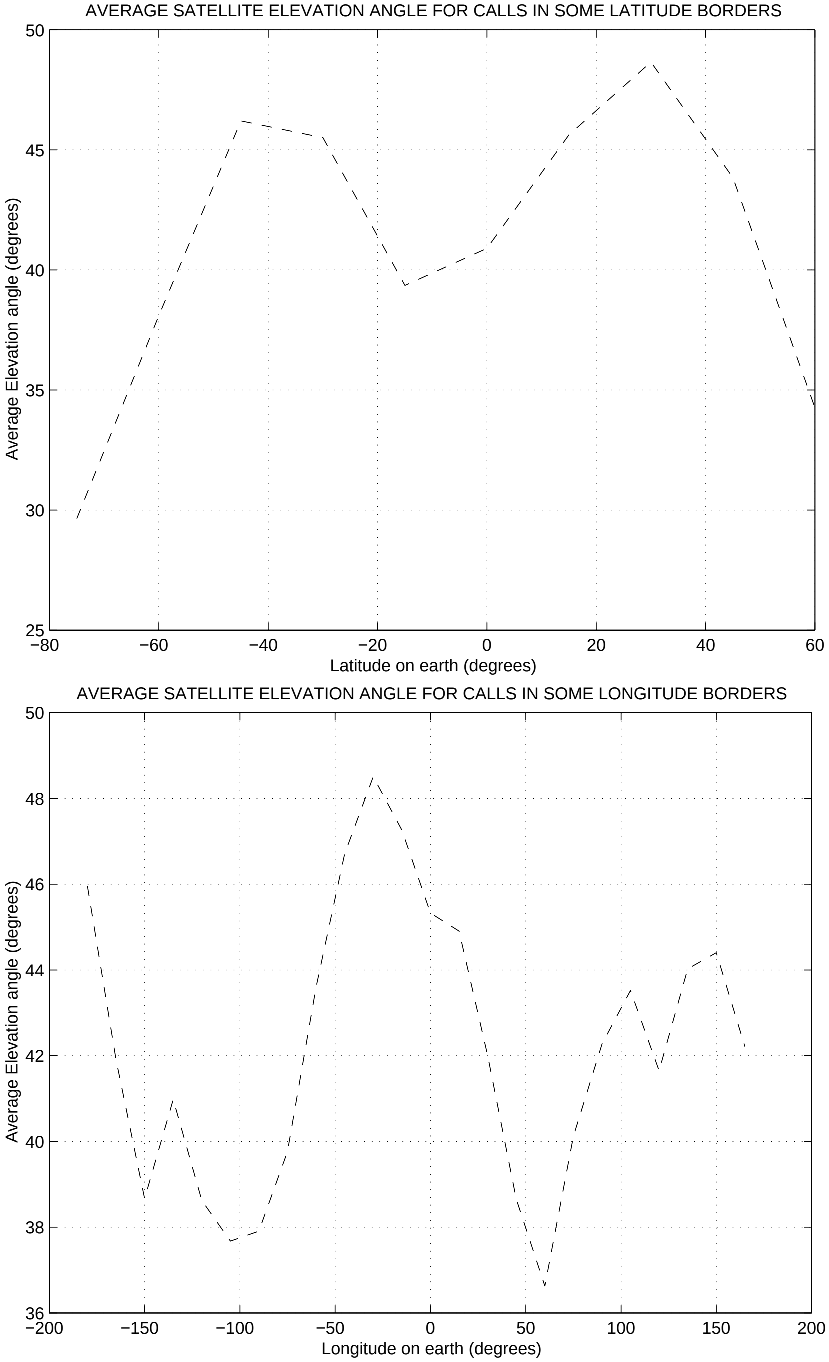

Figure 2.2: Average satellite elevation angle for active calls as a function of latitude and longitude.

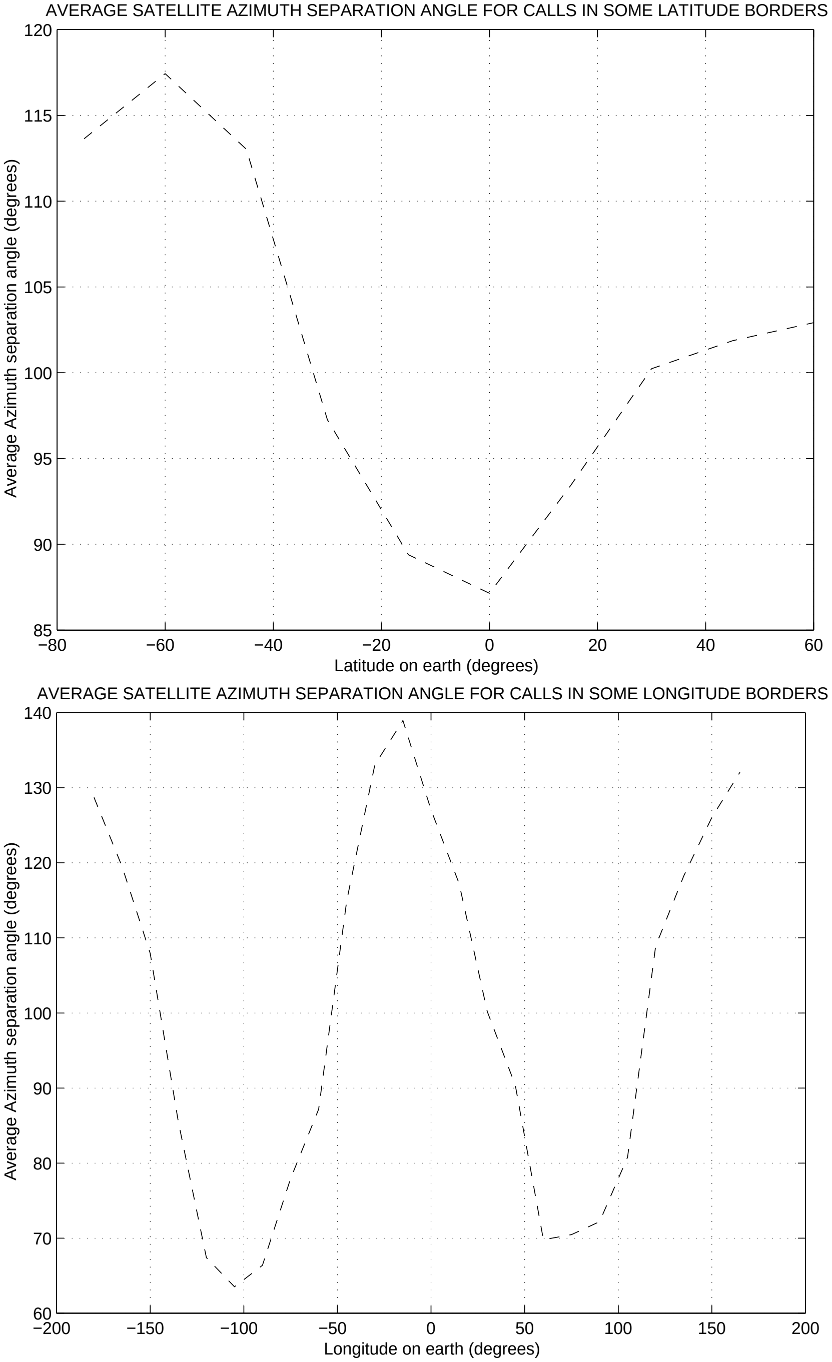

Figure 2.3: Average satellite azimuth separation angle for active calls as a function of latitude and longitude.

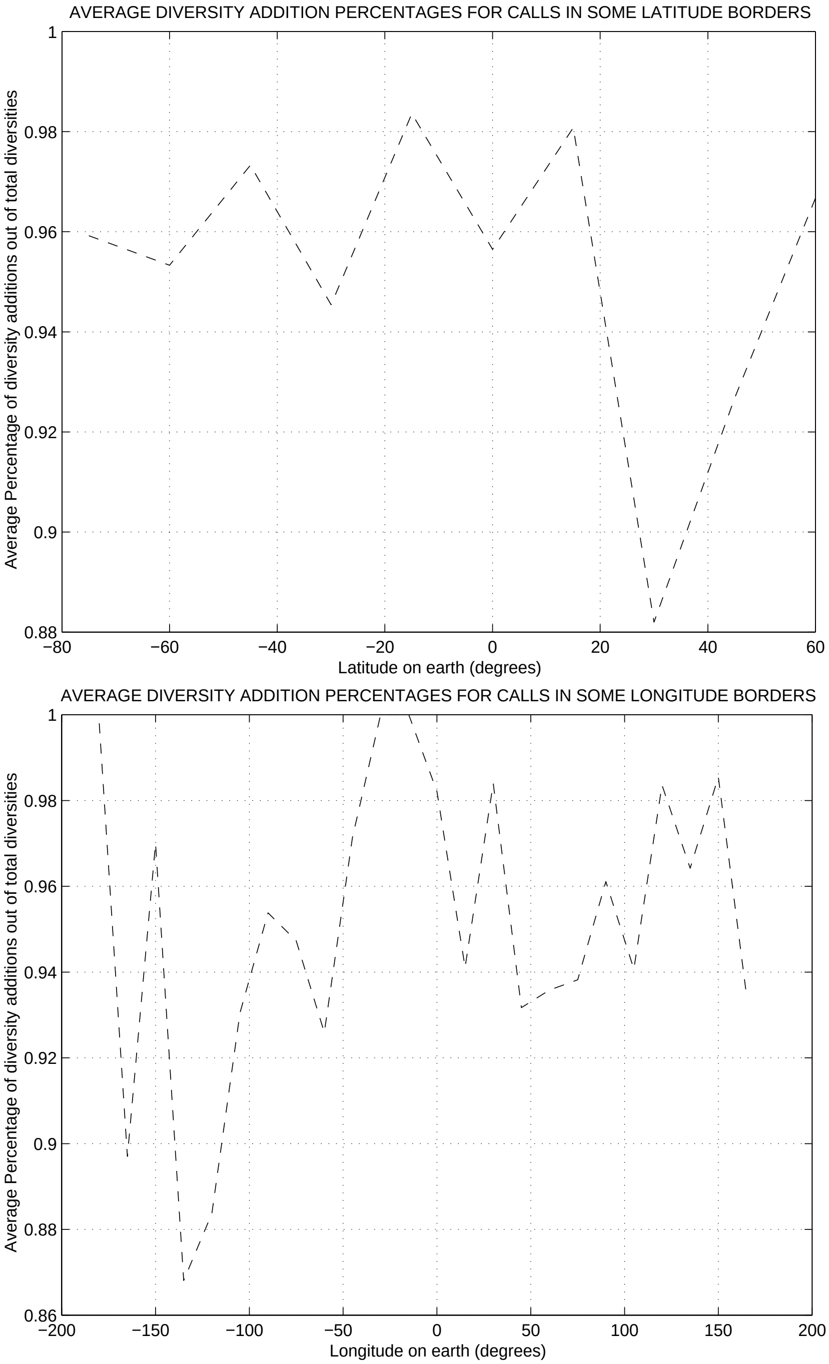

Figure 2.4: Average percentage of *diversity additions* for active calls as a function of latitude and longitude.

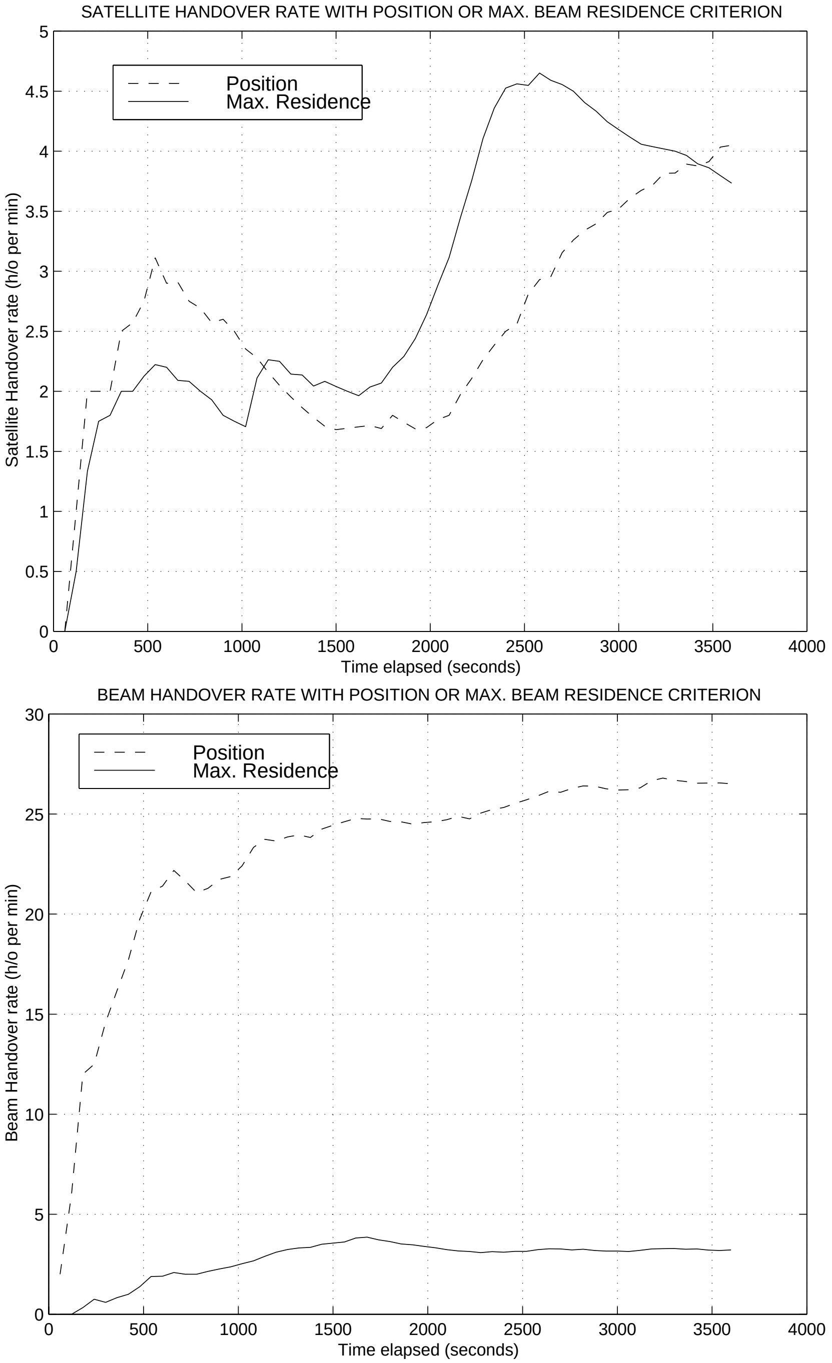

Figure 2.5: Comparison of satellite and beam handover rate under UT position or maximum beam residence time triggered handover event at a region $(-15^o, 0^o)$ longitude , $(15^o, 30^o)$ latitude, with 0.92 calls/sec.

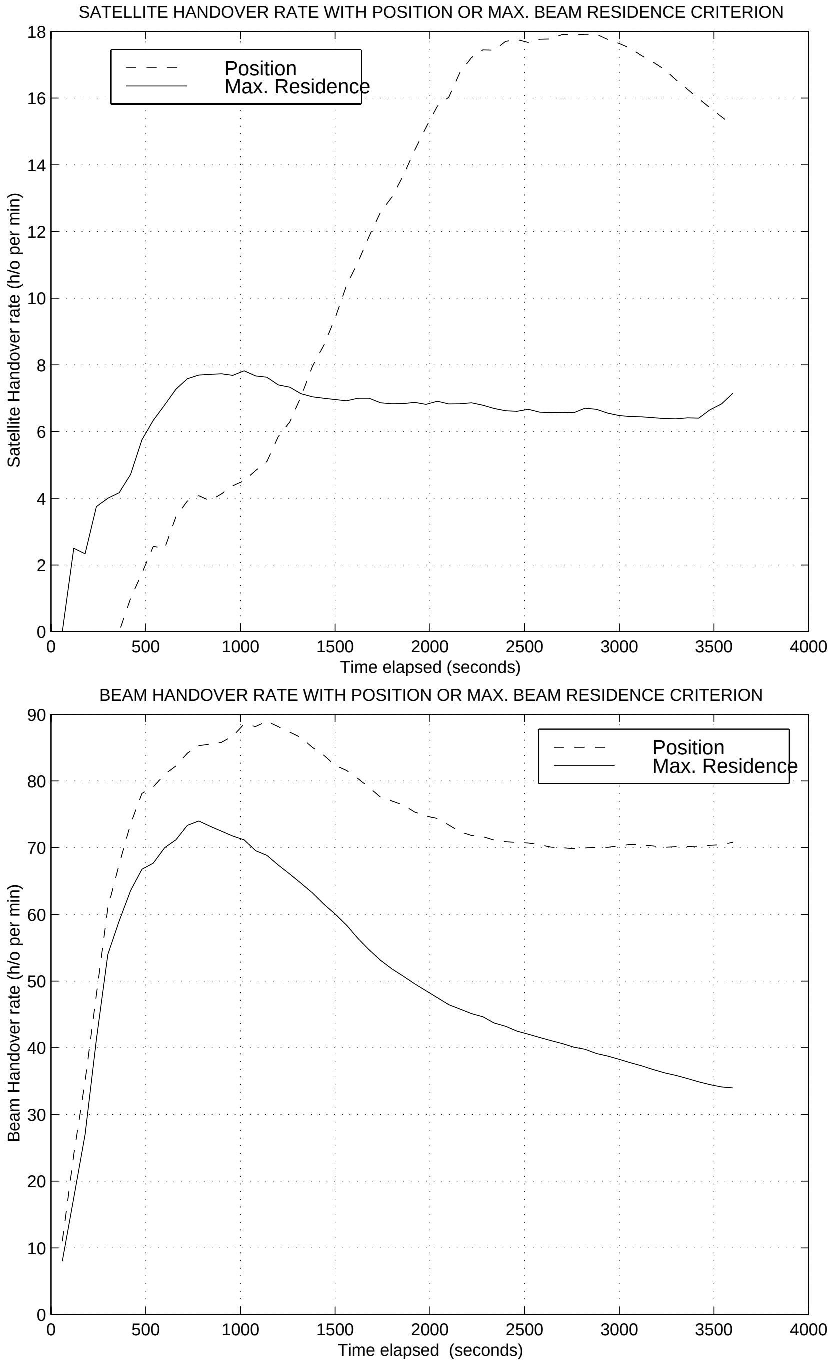

Figure 2.6: Comparison of satellite and beam handover rate under UT position or maximum beam residence time triggered handover event at a region $(60^o, 75^o)$ longitude , $(15^o, 30^o)$ latitude, with 4.87 calls/sec.

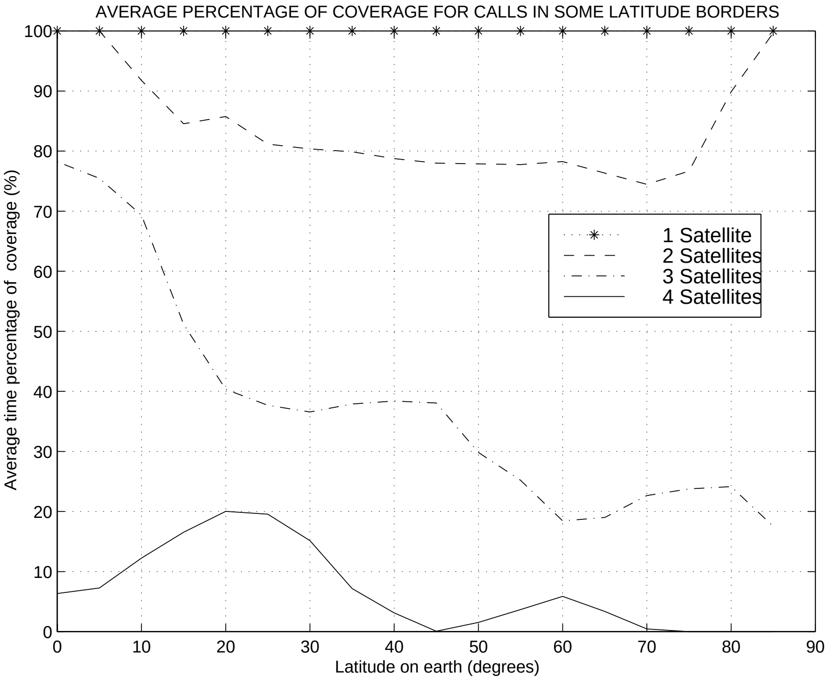

Figure 2.7: Percentage of time coverage from one, two, three or four satellites as a function of latitude.

Chapter 3

A synchronization-based scheme for simultaneous full- and half duplex communication in non-geostationary mobile satellite networks

3.1 Introduction

To provide communication services for mobile or hand-held terminals the critical issues of User Terminal size and cost come into question. To minimize UT mass and cost, the UT must be allowed to operate in half duplex mode on the RF link i.e with non-overlapping transmit and receive times. The option of full-duplex communication between UTs should also exist, for non-hand-held terminals.

Owing to the cost and scarcity of satellite resources, an efficient strategy in allocating them would contribute more to the overall system performance. Moreover, diversity attributes enhance system reliability by providing backup in cases of unpredictable call blockage. The proposed scheme provides a platform for simultaneous full- and half-duplex information transmission, while considerably improving resource management under certain circumstances. It also enables the use of diversity in the case of half-duplex communication in the most efficient way and can be best applicable in GSM-based non-geostationary satellite systems which implicate dynamic resource management and a great number of handoffs and often experience unpredictable changes in signal quality [21]. Results are provided for the case of the ICO mobile satellite system.

When the UT is in diversity mode, it receives from and transmits to the SBS two TCH bursts via two different satellites. The concept of diversity may be considered a waste of resources, since this extra channel that supports the call could be utilized to carry another call. However, an unpredictable blockage during a call can be implemented by a rapid switching from the blocked path to the alternative path with little perceptible effect upon voice quality. A secondary objective of diversity is to allow seamless switching between two alternative voice paths, selecting the path that provides the best combination of voice quality and satellite transmit power.

The option of the type of diversity depends on the type of the UT device. Diversity with non-overlapping reception and transmission intervals is resource-consuming but absolutely consistent with radiation safety standards and is implemented in a class of UTs, destined to be hand-held and support half-duplex communication. Diversity with overlapping reception and transmission intervals can be carried out by UTs supporting full-duplex communication, which are most probably non hand-held.

In section 3.2, basic elements of the structure of the timing and synchronization system of Traffic and Broadcast channels are presented. Section 3.3 provides the motivation of the system we propose, in terms of simultaneous full- and half-duplex diversity connection. Section 3.4 focuses on the delay class feature and suggests two methods for the delay class position determination. Section 3.5 provides more insight into the delay class concept by introducing an assignment method. In section 3.6, several means of achieving UT position determination are given. Finally, in section 3.7 the above methods are quantified.

3.2 Timing and synchronization system

A combination of FDMA/TDMA multiple access scheme is presupposed. FDMA structure instructs that two disjoint frequency bands $[f_1, f_2]$ and $[f'_1, f'_2]$ are assigned to the uplink and the downlink respectively. If B_{RF} is the RF channel spacing, the carrier frequencies for the uplink are

$$f_{c,n} = f_1 + B_{RF} \cdot n, \ 0 \leq n \leq \frac{f_2 - f_1}{B_{RF}} - 1 \tag{3.1}$$

and for the downlink

$$f'_{c,n} = f'_1 + B_{RF} \cdot n, \ 0 \leq n \leq \frac{f'_2 - f'_1}{B_{RF}} - 1 \tag{3.2}$$

The n-th frequency band is $F_n = [f_{c,n}, f_{c,n+1}]$ and $F'_n = [f'_{c,n}, f'_{c,n+1}]$.

Within those n bands, TDMA structure is embedded. In TDMA, users share the same frequency band by accessing the channel in non-overlapping time intervals in a round-robin fashion. Since signals do not overlap, they are orthogonal and the signal of interest is easily extracted by switching the receiver on only during the transmisssion of the desired signal. TDMA can encorporate diversity attributes which make transmission more robust to channel fading.

Assume that the employed TDMA scheme supports N_s TDMA time slots, each of time T. We claim that the minimum (and optimum) number of timeslots per TDMA frame so as the half duplex, diversity enhanced operation and the full duplex operation are supported simultaneously, is $N_s = 6$. Indeed, in order for the half duplex UT transmission requirements to be fulfilled, the UT must transmit and receive at least once in a frame period. Taking into account that a traffic burst is received or transmitted in one timeslot and that reception and transmission should be separated by a small guard time $t_g << T_s$, a minimum of 3 timeslots are required for a single path operation. Therefore, together with the second path, which will establish diversity, 6 timeslots are required.

The timing of Traffic Channels is feasible through the window principle. A reference window is defined as the sequence of three consecutive timeslots. Let the window number be the same as the Absolute Timeslot Number (ATN) of the first timeslot of the window. Windows serve as references at the surface of the earth in order to describe the timing of transmitted and received Traffic Channel (TCH) signals at the UT antenna interface. All windows with the same window number are essentially synchronous everywhere on the earth. The timing of all transmitted and received signals in the system is feasible through a common timing reference, the "system time". The system time is defined by regularly spaced reference instants $\{t_0, t_1, \dots t_n, t_{n+1}, \dots\}$, which coincide with the beginning of a timeslot [22]. By definition, the reference time interval n is the time interval $[t_n, t_{n+1}]$ and the reference window n is the time interval $[t_n, t_{n+3}]$.

The timing of Broadcast Channels (BCCH) is defined according to a periodic cycle of $N_b = 25$ timeslots (frame) at the satellite antenna interface. At the satellite antenna interface, the timeslots of this BCCH frame are aligned with the reference time intervals. The BCCH burst is transmitted in one burst every

BCCH frame. The timeslots used for BCCH transmission are numbered $N_b \cdot n + BCCH\_TN$, where the BCCH timeslot number $BCCH\_TN = 0 \ldots N_b - 1$.

3.3 Motivation and proposed scheme

The basic principle of Time Division and half duplex communication at the UT antenna interface commands that the transmitted and received signals do not overlap. The basic requirement to be fulfilled is that the transmitted and received signals by a UT have to be accomodated in a given reference window. This scheme allows for diversity when the following holds:
Fact:

The (GPS) Window number corresponding to one pair
of forward/reverse link TCH bursts at the UT
=(Window number for the second diversity link + $N_s/2$*) mod* N_s

In general, the two physical links are allocated in two cells belonging to two different satellites and timing of signals is controlled independently in the two windows.

In some cases, the transmit and receive time intervals are arranged symmetrically within a reference window. However, the actual relative position of the transmitted and received bursts at a UT antenna interface and their position within a reference window depend on the propagation delay T_p between the satellite and the UT. For different UTs the time delay varies within a certain range, $T_0 - \delta T_p \leq T_p \leq T_0 + \delta T_p$, around a nominal delay value T_0, which obviously corresponds to the symmetrical arrangement. A small guard time t_g accounts for transmit-receive switching, local oscillator re-tuning and residual

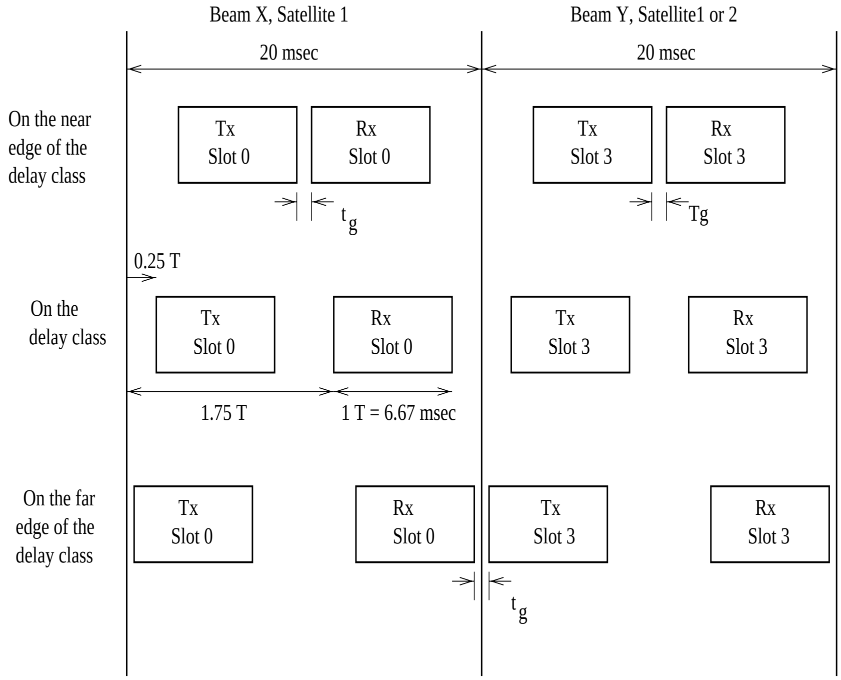

Figure 3.1: Transmit-Receive traffic intervals in the case of diversity

timing errors and should be of the order of μsec. Generalizing the argument, we can say that for any transmission-reception configuration scheme with guard time t_g and timeslot duration T_s:

$$\delta T_p = 0.25T - \frac{t_g}{2} \tag{3.3}$$

Then, transmission and reception will begin at times

$$T_{TX} = 6.25T + T_p - T_0\,(mod 6) \tag{3.4}$$

and

$$T_{RX} = 1.75T + T_p - T_0 \tag{3.5}$$

Within a satellite footprint, time delay ranges between two extremes. To satisfy the above requirements for all UTs irrespective of their position, many nominal values $T_{0,i}$ must be defined. Each of them comprises a class of traffic channels C_i , whose delay lies in the range:

$$T_{0,i} - \delta T_p \leq T_p \leq T_{0,i} + \delta T_p \tag{3.6}$$

System time is distributed from the SBS to the UTs via Control Channels. The burst timing is pre-corrected by the SBSs so as to leave the satellite at reference time in all cells. Thus, all channels belonging to the same delay class are essentially aligned in the satellite antenna interface. The UT derives its own timing reference from the signal received from the network, which will be delayed by the propagation time between the satellite and the UT, T_p. Pictorially, each nominal delay value corresponds to a contour on the surface of the earth, centered in the subsatellite point and can be referred to as *Delay Class.*

The next issue that comes into question is what is the optimum (fixed) number κ of constant delay circles that should be defined. By adopting the Delay Class concept under the constraint of combined full- and half-duplex communication capabilities, channel configuration can be redefined. A set of frequency bands is assigned to each beam, which can be reused by another beam, provided that cochannel interference constraints are not violated. Because the locations of the contours within the satellite footprint on the surface of the earth have a certain position with respect to the beam pattern in the footprint, every Class serves a particular set of beams.

3.4 Derivation of the exact position of Delay Classes

The derivation of the exact position of the delay class contours on the earth surface can be realized by utilizing either the Load Balancing concept or the requiremnt of an efficient Delay Class handover scheme with a big handover time margin.

3.4.1 The Load Balancing concept

Resources within a satellite footprint are assumed to belong to a pool and are distributed on a per beam basis. Channels are assigned with Dynamic Channel Allocation (DCA) schemes to account for non-geostationary satellite movement and traffic variations. However, of particular importance is the fact that beams obtain different shapes when projected on the earth surface due to the earth curvature. Thus, beams close to the subsatellite point remain almost of circular shape, whereas beams close to the footprint edge become rather elliptical and elongated. For uniformly distributed traffic on the footprint area, the expected number of calls served by outmost beams is greater because of their coverage area. Despite DCA, this phenomenon may cause problems in QoS level and efficient resource management. The Delay Class concept can be used as a supplement to compensate for that by applying in essence the Load Balancing principle to balance the distribution of mobiles along the footprint.

Suppose that a fixed number, κ, of Delay Classes has been defined. The total number of Delay Classes will depend on the length of the overlap area between the adjacent Delay Classes, which, in turn, reflects the waiting time tolerance for a call in order to be assigned to either class of channels. Assuming that all delay Class handover requests concerning mobiles located in overlap areas between two delay classes are queued, a relatively big overlap area allows enough time tolerence for a delay class transition decision. The determination of the time delays $T_{o,i}$ corresponding to the delay classes (and therefore their position) is accomplished using the *Load Balancing* principle as guideline. In the following, a simple heuristic algorithm for the derivation of Delay Class positions is provided.

The idea of this algorithm is to ensure that the specific positions of the delay classes divide the earth surface into zones with equal area, so as to be consistent with traffic balancing on the earth surface, assuming uniform spatial traffic distribution. We begin by defining a "cup" of height h_0 on the earth sphere, with area A/κ, where A is the area of hemisphere and κ is the (predefined) number of delay classes. We proceed by finding at each iteration the height h_j defining zone j, based on the equal zone area principle. The outcome of the algorithm is a sequence of time delay values $\{T_{o,n}\}_{n=1}^{\kappa}$, denoting the exact position of each delay class.

Algorithm A : Derivation of delay class positions

1. Define the minimum elevation angle θ_{min} for an acceptable connection path.

2. Derive the corresponding (maximum) central angle ϕ_{max} by using the general formula:

$$\phi(\theta) = \frac{\pi}{2} - \sin^{-1}\left(\frac{R_E}{R_E + H}\cos\theta\right) - \theta \qquad (3.7)$$

3. For iteration $j = 0$ do steps 4-8.

4. Evaluate the height $h_0 = R_E\,(1 - \cos\theta_{min})$ (figure 3.2).

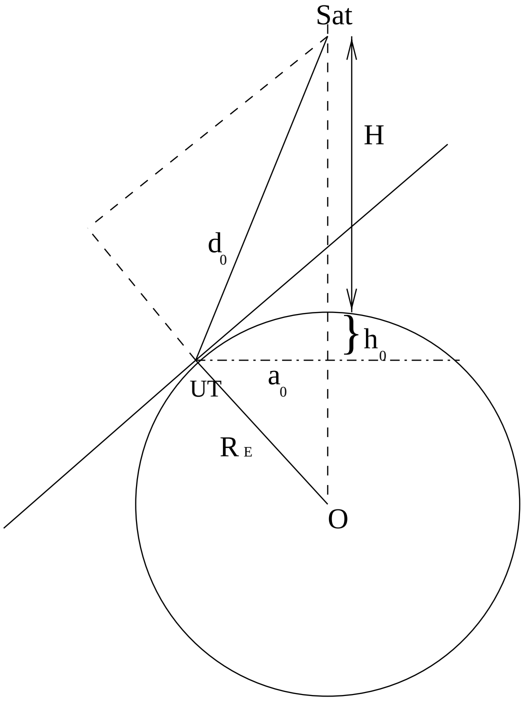

Figure 3.2: Illustration of the first step of the algorithm for delay class delay computation (Algorithm A).

5. Find the area of the "cup" $A = 2\pi R_E h_0$

6. Compute the distances

$$a_0^2 = R_E^2 - (R_E - h_0)^2 \tag{3.8}$$

$$d_0^2 = a_0^2 + (H + h_0)^2 = H^2 + 2(R_E + H)h_0 \tag{3.9}$$

7. Find the corresponding time delay to the satellite

$$\tau_0 = \frac{d_0}{c} \tag{3.10}$$

8. For the j-th iteration do steps 9-12

9. Select constant height

$$h_j = A/(2\pi\kappa R_E) \tag{3.11}$$

10. Compute the distance

$$a_j^2 = R_E^2 - \left(R_E - \sum_{i=0}^{j} h_i \right)^2 \tag{3.12}$$

11. Find the distance

$$d_j^2 = a_j^2 + \left(H + \sum_{i=0}^{j} h_i \right)^2 = H^2 + 2\left(R_E + H\right) \sum_{i=0}^{j} h_i \tag{3.13}$$

12. Find the corresponding time delay

$$\tau_j = \frac{d_j}{c} \tag{3.14}$$

3.4.2 Increased Handover time margin

The Delay Class system may be designed by having the handover time margin as a performance measure. Delay class contours should be placed in such positions, so that the delay class overlap regions are of a certain length, according to system specifications. A big overlap region amounts to a greater time tolerance for delay class handover, since the time margin in which the presumed transition must take place is greater. Different treatment should also be applied for delay classes belonging to beams near the subsatellite point than those serving edge beams, the reason being that delay changes with a higher rate in the latter case.

Consider the maximum rate in which time delay changes, owing to the relative movement of the mobile and the satellite as $(dt_d/dt)_{max}$, (measured in $\mu sec/sec$) as it appears in edge beams. Assume also that the time margin associated with a reliable delay class transition is set to T_{margin}, depending upon several system parameters. Then the overlap region between any two delay classes, $T_{overlap}$, should satisfy

$$T_{overlap} \geq \left(\frac{dt_d}{dt} \right)_{max} \times T_{margin} \tag{3.15}$$

3.5 Delay Class determination for Resource allocation

3.5.1 Problem Statement

In order for a new channel to be assigned to a call that will experience handover or will be given diversity, the Handover Management (HOM) Software Unit of the Land Satellite Resource Management System (LSRMS) must request the appropriate resources from the Dedicated Channel Management (DChM) Software Unit, while at the same time providing the satellite, beam and delay class to it. DChM will then be able to allocate adequate resources to the call.

If the serving satellite remains the same after the handover event, i.e a beam or delay class Handover occurs, the new delay class is known because the SBS maintains the timing to the current satellite. However, in the event of a satellite handover, synchronization is lost, owing to the independence of satellite synchronization systems. The derivation of the delay class in the new satellite is the cornerstone in continuing to keep track of the timing of the system. To this end, the time delay from UT position to the new satellite is not available and has to be computed by the HOM at the SBS.

Two methods can be used by the SBS to determine this delay:

- METHOD 1: The SBS can use the current available known estimate of the UT position together with the satellite ephemeris data, to derive an estimate of the delay.
- METHOD 2: The SBS can request a one-shot measurement report from the UT. In that mode, the SBS provides the UT with the BCCH timeslot,

nominal frequency and rough propagation delay information with respect to the new satellite. The UT searches for and measures the new BCCH relative to the serving satellite and reports the difference between the actual and rough propagation delay, $T_p - \tilde{T_p}$ with respect to the new satellite back to the SBS, which can now determine the new propagation delay with high accuracy.

The first method is easier and faster to implement while the second one requires additional bandwidth on the air interface for the transmission of messages and is also time-consuming, but provides more accuracy. Therefore, the first should be given priority and used whenever the known UT position is accurate enough to provide a reasonable estimate of the delay and the second alternative should be used in the opposite case.

3.5.2 Computation of the Tolerance Region

Each UT is characterized by a unique actual location on the earth and a unique time delay to a satellite. However, only estimates of the above quantities are available to the system. Estimated position is also referred to as known position.

The case of a beam with two delay classes is considered (figure 3.3). Depending on the path delay value from the UT to the satellite, the UT can be classified in one of the three regions. Because of the inaccuracy in the delay evaluation, the UT may seem to reside in a different region from its actual one. More explicitly:

- If the UT's actual position is in Region 1, then the wrong delay class is

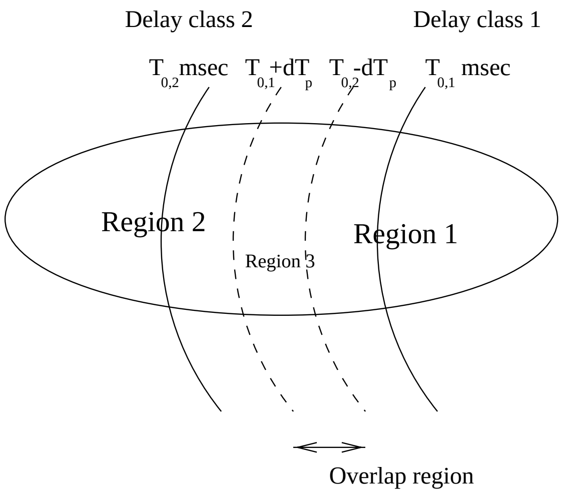

Figure 3.3: Demonstration of the situation in a two–delay class spot beam with the corresponding overlap region.

assigned either if the known UT position is in Region 2 or if the known UT position is in Region 3 and delay class 2 is selected.

- If the UT's actual position is in Region 2, then the wrong delay class is assigned, either if the known UT position is in Region 1 or if the known UT position is in Region 3 and delay class 1 is selected.

- If the UT's actual position is in Region 3, then either of the two delay classes can serve the call. In the worst case there will be two successive delay class handovers without undesirable consequences to the connection of the ongoing call.

Therefore, a real corruption of the call due to inappropriate delay class assignment occurs only when the difference between the actual and the known position corresponds to a difference in the time path delay *greater than the length of the overlap region.*

The correct delay class must be allocated to the call, so that the probability of an erroneous resource assignment is minimized. Suppose that the destination beam has κ_i delay classes. The following procedure provides a heuristic in determining the maximum allowed error in the UT position determination.

Algorithm B: tolerance computation

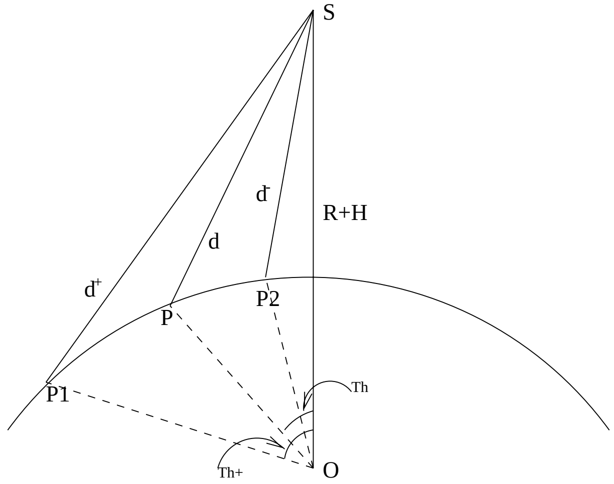

Figure 3.4: Algorithm A for tolerance region computation

1. Divide the set of beams in the satellite footprint into subsets $B_1, B_2, \ldots B_n$ such that every beam in B_j has κ_j delay classes. All beams belonging to subset B_j form a toroid. Subset B_n comprises essentially all outmost beams and subset B_1 includes only the nadir beam.

2. For each delay class i, serving beams of B_j find the time lengths of the overlap regions:

$$x_{i,j} = T_{0,i} - T_{0,(i+1)} - 2|\delta T_p| \,, \; T_{0,i} > T_{0,(i+1)} \tag{3.16}$$

3. For each subset B_j select

$$x_j = \min_{i=1\ldots\kappa_j} x_{i,j} \tag{3.17}$$

to account for the worst case scenario (the delay class with the smallest overlap region). This is the maximum allowed tolerance between the actual and the estimated time delay $\hat{e}_{max,j} = max|T_p - \tilde{T}_p|$ for mobiles belonging in beams of subset B_j.

4. Evaluate the corresponding tolerance in the UT-satellite path distance

$$\Delta d_j = c\hat{e}_{max,j} \tag{3.18}$$

5. Consider the two "extreme" paths

$$d_{i_j}^{\pm} = cT_{0,i} \pm \Delta d_j, \;\; i = 1\ldots\kappa_j \tag{3.19}$$

and the corresponding central angles

$$\theta_{i,j}^{\pm} = \frac{R_E^2 + (R_E + H)^2 - \left(d_{i,j}^{\pm}\right)^2}{2R_E\left(R_E + H\right)} \tag{3.20}$$

6. Compute the radius of each circular tolerance region as

$$L_{i,j} = \frac{R_E}{2}\left|\theta_{i,j}^{+} - \theta_{i,j}^{-}\right| \tag{3.21}$$

under the assumption of "flat earth".

7. Select

$$L_j = \min_{i=1\ldots\kappa_j} L_{i,j} \tag{3.22}$$

This denotes the maximum allowed error in the UT position determination as will be described in the following.

3.5.3 Derivation of the delay class assignment method

The following procedure describes the selection of the appropriate delay class assignment method, which essentially is the reliable estimation of UT position. The fact of crucial importance is whether the mobile's known position is located in the overlap region between two delay classes or not. If the mobile is not found to reside in an overlap region, but can be unambiguously assigned to one delay class, the only situation when there could be an incorrect assignment, would be if the UT was near the border of regions 1 and 3 (point P, in figure 3.3) or near the border of regions 2 and 3 (point Q, in figure 3.3) and the estimation error was greater than the position error corresponding to the overlap region.

1. Execute Algorithm B above.

2. For each UT in the satellite footprint do steps 3-6.

3. Compute the actual time delay and Doppler.

4. Randomly generate delay and frequency values normally distributed and estimate UT position.

5. Compute the earth great arc distance between the estimated and the true UT position and determine where the known (estimated) UT position is located.

6. Determine also the exact beam and beam subset B_j where the UT is located. Thus the number of delay classes assosiated with this subset is known.

    - CASE A1 :Known position is in Region 1 or 2 and $|UT_{act} - UT_{known}| < L_j$. Method 1 should be used to find the delay.

- CASE A2 :Known position is in Region 1 or 2 and $|UT_{act} - UT_{known}| \geq L_j$. Method 2 should be used to find the delay.
- CASE B1 :Known position is in Region 3 and $|UT_{act} - UT_{known}| < \min\{\tau_1, \tau_2\}$. Method 1 should be used to find the delay.
- CASE B2 :Known position is in Region 3 and $|UT_{act} - UT_{known}| \geq \min\{\tau_1, \tau_2\}$. Method 2 should be used to find the delay.
- Use the delay value to assign the call ultimately to a Delay Class

It should be mentioned that:

- In the case when the UT is located in an overlap area, the parameters

$$\tau_1 = T_{0,i+1} + \delta T_p - \tilde{T}_p \tag{3.23}$$

$$\tau_2 = \tilde{T}_p - (T_{0,i} - \delta T_p) \tag{3.24}$$

  are computed and the UT is supposed to be in the overlap region of delay classes i and $i+1$, with $T_{0,i} > T_{0,(i+1)}$.

- The property

$$|UT_{act} - UT_{known}| > \min\{\tau_1, \tau_2\} \tag{3.25}$$

  means that the error region covers Region 3 and part of 2 and/or 1.

By this algorithm the probability of an incorrect delay class assignment is exactly zero and the resource assignment is realized in the most efficient way in terms of time delay and signalling overhead in the network.

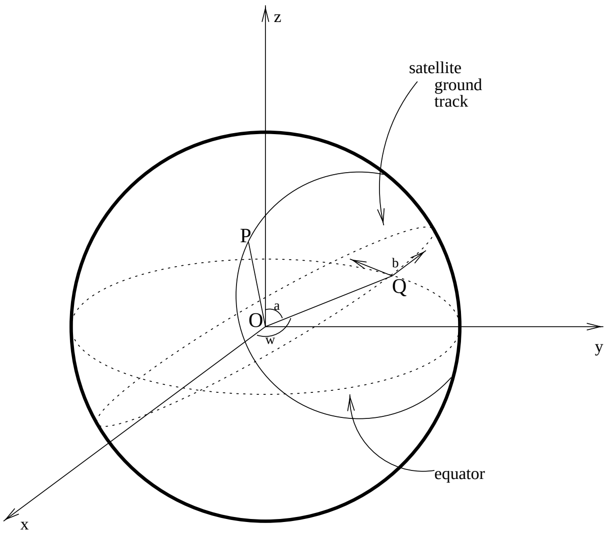

Figure 3.5: Angles computed for UT position determination

3.6 UT position determination

Assume that the vectors $\vec{P}(t)$ and $\vec{S}(t)$ represent the UT and the satellite position and that Q is the vector (point) corresponding to the subsatellite point. Define the following angles (figure 3.5)

- The *distance angle* a of the UT from the subsatellite point, reflecting UT distance from satellite nadir.

- The *azimuth angle* b of a UT at the subsatellite point relative to the satellite motion direction.

- The *true anomaly* w, denoting the position of a satellite into its orbit, given by:

$$w(t) = \frac{2\pi t}{T} \tag{3.26}$$

where T is the period of the satellite.

3.6.1 UT position determination using Time delay and frequency Doppler calculations

Denote by $\vec{D}(t) = \vec{S}(t) - \vec{P}(t)$ the distance vector from the UT to the satellite, t_d the time delay from the UT to the satellite, f_d the Doppler frequency because of the relative movement and f the base frequency of the signal. Then the path delay and Doppler offset for a radio signal traveling between the UT and the satellite are respectively [4]

$$t_d = \frac{|\vec{D}(t)|}{c} = \frac{|\vec{S}(t) - \vec{P}(t)|}{c} \tag{3.27}$$

$$f_d = -\frac{\vec{D}'(t)}{c} f \tag{3.28}$$

where

$$\vec{D}'(t) = \frac{d|\vec{D}(t)|}{dt} = \frac{\vec{D}(t)}{|\vec{D}(t)|} \frac{d\vec{D}(t)}{dt} = \frac{\vec{S}(t) - \vec{P}(t)}{|\vec{S}(t) - \vec{P}(t)|} \left[\vec{S}'(t) - \vec{P}'(t)\right] \tag{3.29}$$

Therefore, given a, b and w, the cartesian coordinates can be computed and the above equations can be used to derive the time delay and Doppler. It is more straightforward, however, to calculate the time delay by using the law of cosine, as:

$$t_d = \frac{\sqrt{R_E^2 + (R_E + H)^2 - 2R_E(R_E + H)\cos a}}{c} \tag{3.30}$$

To calculate the Doppler, we convert the representation of $\vec{P}(t)$ from a, b and w to cartesian coordinates. This conversion can be done in three steps:

1. Assuming that the subsatellite point Q is on the equator, on the Greenwich meridian, and that the satellite orbit is in the $x-y$ plane so that the ground

track concides with the equator, compute the x, y, z coordinates of $\vec{P}$.

$$\begin{aligned} x_{p,1} &= R_E \cos a \\ y_{p,1} &= R_E \sin a \cos b \\ z_{p,1} &= R_E \sin a \sin b \end{aligned} \tag{3.31}$$

2. Rotate $\vec{P}$ and $\vec{Q}$ with respect to the z-axis, so that Q has an angle w with respect to the x axis.

$$\begin{aligned} x_{p,2} &= x_{p,1} \cos w - y_{p,1} \sin w \\ y_{p,2} &= x_{p,1} \sin w + y_{p,1} \cos w \\ z_{p,2} &= z_{p,1} \end{aligned} \tag{3.32}$$

3. Assuming that the satellite orbit has an inclination angle of i , rotate $\vec{P}$ and $\vec{Q}$ with respect to the x-axis by i degrees.

$$\begin{aligned} x_{p} &= x_{p,2} \\ y_{p} &= y_{p,2} \cos i - z_{p,2} \sin i \\ z_{p} &= y_{p,2} \sin i + z_{p,2} \cos i \end{aligned} \tag{3.33}$$

Combining the above equations, we write $\vec{P}$ as:

$$\vec{P} = \begin{bmatrix} x_p \\ y_p \\ z_p \end{bmatrix} = R_E \begin{bmatrix} \cos a \cos w - \sin a \cos b \sin w \\ \cos a \sin w \cos i + \sin a \cos b \cos w cosi - \sin a \sin b \sin i \\ \cos a \sin w \sin i + \sin a \cos b \cos w \sin i + \sin a \sin b \cos i \end{bmatrix} \tag{3.34}$$

The motion of the mobile can be written as

$$\vec{P\prime} = \frac{2\pi}{T} \begin{bmatrix} -\cos a \sin w \cos i - \sin a \cos b \cos w cosi + \sin a \sin b sini \\ \cos a \cos w - \sin a \cos b \sin w \\ 0 \end{bmatrix} \tag{3.35}$$

The point $\vec{Q}$ can be considered to be a special point $\vec{P}$ with $a = 0$. The vector $\vec{S}$ is an extension of $\vec{Q}$ and can be expressed in terms of a, b, and c as:

$$\vec{S} = (R_E + H) \begin{bmatrix} \cos w \\ \sin wcosi \\ \sin w \sin i \end{bmatrix} \tag{3.36}$$

and the motion of the satellite is

$$\vec{S'} = \frac{2\pi(R_E + H)}{T} \begin{bmatrix} -\sin w \\ \cos wcosi \\ \cos w \sin i \end{bmatrix} \tag{3.37}$$

With some algebraic manipulations we can show that

$$(\vec{S} - \vec{P})(\vec{S'} - \vec{P'}) = 2\pi(R_E + H)\sin a \left(\frac{\cos b \cos i - \cos wsinb \sin i}{T} - \frac{\cos b}{T} \right) \tag{3.38}$$

Therefore, the Doppler can be computed as

$$F_d = -\frac{2\pi R_E(R_E + H) f \sin a}{c\sqrt{R^2 + (R_E + H)^2 - 2R_E(R_E + H)\cos a}} \left(\frac{\cos b \cos i - \cos wsinb \sin i}{T} - \frac{\cos b}{T} \right) \tag{3.39}$$

Given the satellite position $\vec{S}$, the measured time delay t_d and the measured Doppler F_d, the position $\vec{P}$ of the UT can be determined as follows:

1. Calculate angle a by the equation

$$a = \cos^{-1} \frac{R_E^2 + (R_E + H)^2 - (ct_d)^2}{2R(R + H)} \tag{3.40}$$

2. Solve the equation 3.39 for b

3. Calculate w from equation 3.36

4. Convert the (a, b, w) representation of $\vec{P}$ to desired form (e.g. latitude and longitude)

3.6.2 Other means of determining UT position

In the previous section, the procedure of finding the UT position by using the time delay and Doppler from one satellite has been explicitly presented. Additional accuracy is achieved by using the measurements of two satellites. Specifically the following methods are used:

- METHOD I: Delay and Doppler from one satellite.
- METHOD II: One time delay and the differential delay measurement of two satellites.
- METHOD III: One frequency and the differential delay measurement of two satellites.
- METHOD IV: One time delay and the differential Doppler measurement of two satellites.
- METHOD V: One frequency and the differential Doppler measurement of two satellites.
- METHOD VI: The differential time delay and the differential Doppler measurement of two satellites.

3.7 Simulations and results

To illustrate the performance of the UT position determination algorithm and its impact on the correct delay class selection, experiments have been executed for the ICO satellite system with signal base frequency $f = 2.01\, GHz$.

We found that $max_j \kappa_j = 3$ in order to be consistent with the requirements of the previous section. Therefore by using Algorithm A, the tolerance range for the subset of beams with two and three delay classes were found to be $142km$ and $24km$ respectively. Beams with one delay class are of no interest, since they exhibit no delay class transitions. Therefore the results depicted in Table 3.1 can be used in the determination of percentage of times when each one of the methods for the correct delay class determination can be used. If UT position error is less than L_j for $y_j\%$ of the times for some j, Method 1 can be used to determine the new delay class for $y_j\%$ of the times and Method 2 should be used limitedly, for the rest $(100 - y_j)\%$ of times. In particular, we can deduce the following:

- Method 1 has excellent performance if the new beam has two delay classes, irrespective of the utilized measurements. In that case it can be used almost entirely $(97\% - 98\%)$. Therefore considering measurements from two satellites is not essential and can be avoided to reduce signalling overhead.

- If the new beam has three delay classes and measurements from one satellite are used, Method 1 can be characterized relatively reliable, because it attains fairly satisfactory results for $80\% - 83\%$ of the times, irrespective of the measurement method. Method 2 can be utilized for $17\% - 20\%$ of the times, depending on the measurement method.

- For the case of a three delay class beam, high performance can be reached by Method 1 if differential delay measurements between two satellites are used together with differential Doppler or time delay from one satellite. In that case, Method 1 can be used almost exclusively. However, method

1 exhibits moderate performance (64%) in a three delay class beam when one frequency and the differential Doppler measurement from two satellites are used.

- If the UT's known position lies in the overlap region, Method 2 can be used when the radius of the error region exceeds certain threshold values, depending entirely on the known position. The percentage of times when it is used is dependent on the instantaneous UT known position and time delay can easily be calculated in a similar fashion.

To back up the analysis for handover described in previous sections, handover rates in several locations on the earth, as well as on the entire surface have been recorded. From figures 3.6- 3.9, we focus on particular locations on earth and derive an approximation of the expected number of delay class handover rate. It becomes evident that, after a transient period of about 2500 sec, beam handover rates tend to stabilize (the number of beam handovers increases in an almost linear fashion). Satellite and delay class handovers are subject to stabilizaton up to a certain extend. Certainly, the number of transitions depends decisively on the traffic induced on that particular location. In steady state, we observed that about 3 and 30 delay class transitions per minute occur, when the traffic is 1.69 calls/sec and 4.87 calls/sec respectively, a fact in favor of the conclusion that under heavy traffic load, the number of transitions increases rapidly. Delay class handover rate becomes in a way predictable, as the transition will occur only within beams with more than one delay class. These kinds of graphs are of great importance in resource planning and forecasting in different times of the day. Similar graphs can be obtained for every hour of the day, given the expected amount of traffic and its variations.

Of interest is a comparison of the situation that arises if we adopt the maximum beam residence handover criterion. For a fixed square location on earth, it was observed that beam handover rate was reduced from 25 handovers per minute to 4 handovers per minute. Delay class handover increased in that case as expected, from $2-3$ transitions per minute up to $8-10$ transitions, a fact rather expected, since the UT resides in the beam location area the maximum time and experiences all delay class transitions. In general, the grade at which the delay class handover rate increases depends on whether the UT is located on edge beams with more than one delay classes. Satellite handover rates remain mostly unaffected after adoption of maximum beam residence criterion. From figure 3.12 we obtain the order of magnitude of each of the aforementioned kinds of transition for the entire earth.

Table 3.1: Percentage of times when the UT lies within a certain range for different UT position determination methods.

Method	in < 25km	in < 50km	in <100km	in < 150km
Method I	82%	92%	96%	97%
Method II	90%	95%	97%	98%
Method III	84%	95%	98%	99%
Method IV	85%	94%	97%	99%
Method V	64%	93%	99%	> 99%
Method VI	97%	93%	> 99%	> 99%

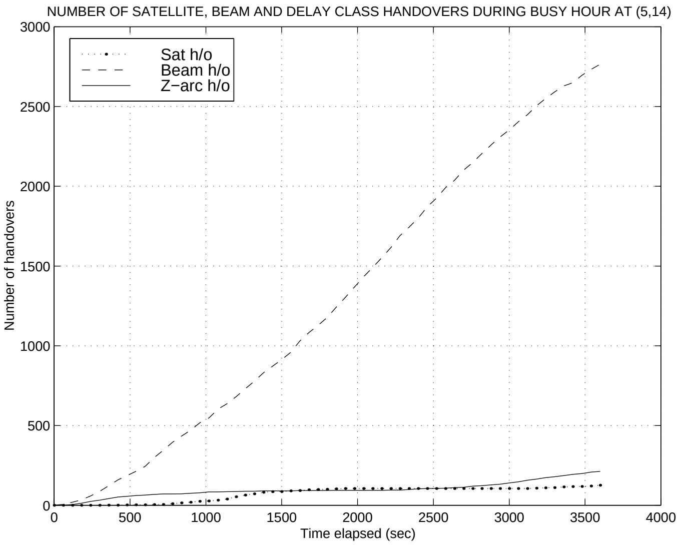

Figure 3.6: Satellite, beam and delay class handovers at a region $(15^o, 30^o)$ longitude, $(15^o, 30^o)$ latitude, with 1.69 calls/sec .

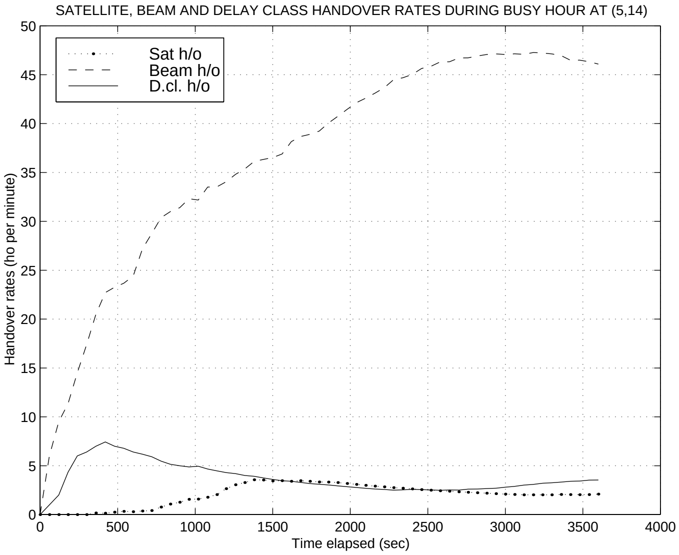

Figure 3.7: Satellite, beam and delay class handover rates at a region $(15^o, 30^o)$ longitude , $(15^o, 30^o)$ latitude, with 1.69 calls/sec.

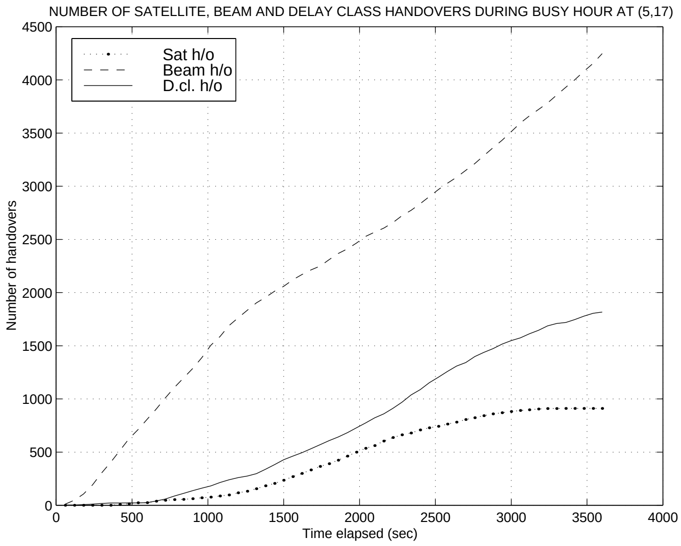

Figure 3.8: Satellite, beam and delay class handovers at a region $(60^o, 75^o)$ longitude, $(15^o, 30^o)$ latitude, with 4.87 calls/sec.

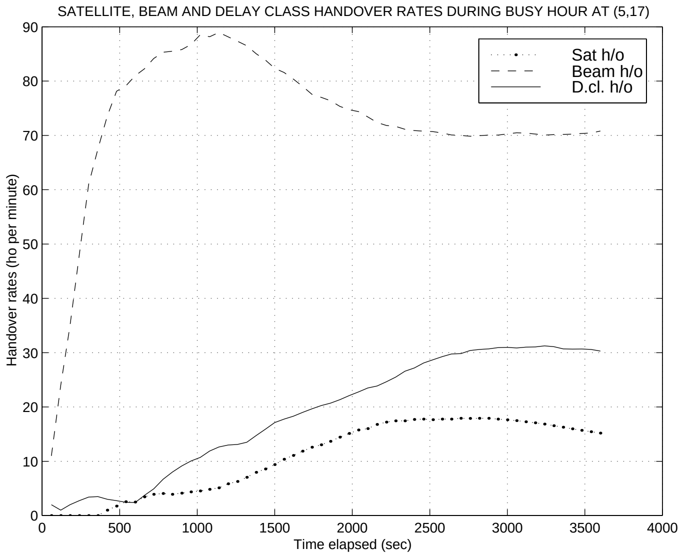

Figure 3.9: Satellite, beam and delay class handover rates at a region $(60^o, 75^o)$ longitude, $(15^o, 30^o)$ latitude, with 4.87 calls/sec.

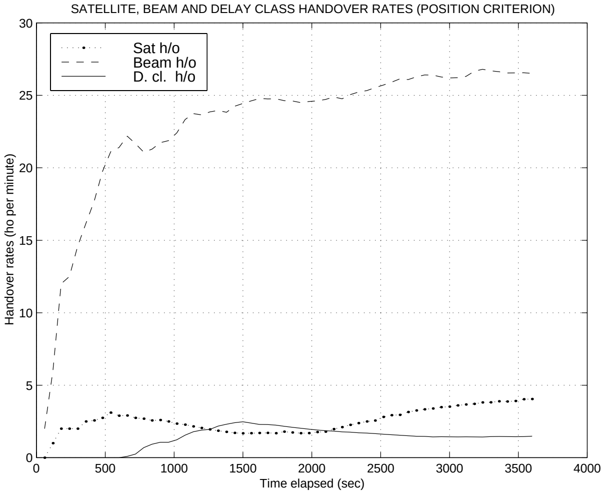

Figure 3.10: Satellite, beam and delay class handover rates at a region $(-15^o, 0^o)$ longitude, $(15^o, 30^o)$ latitude, with 0.92 calls/sec, by using the UT position as the triggering event for handover.

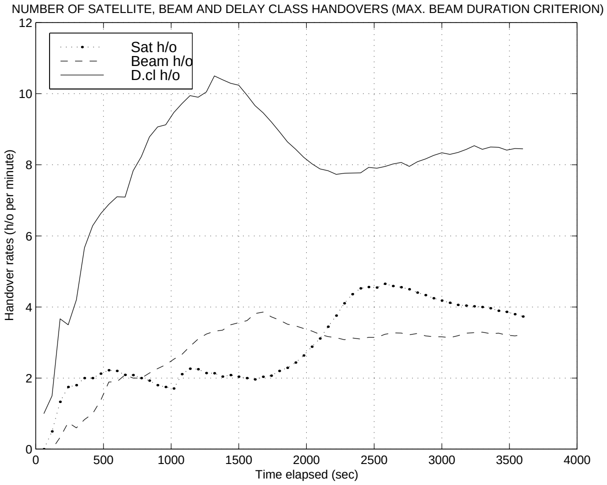

Figure 3.11: Satellite, beam and delay class handover rates at a region $(-15^o, 0^o)$ longitude, $(15^o, 30^o)$ latitude, with 0.92 calls/sec, by using the maximum beam residence time as the triggering event for handover.

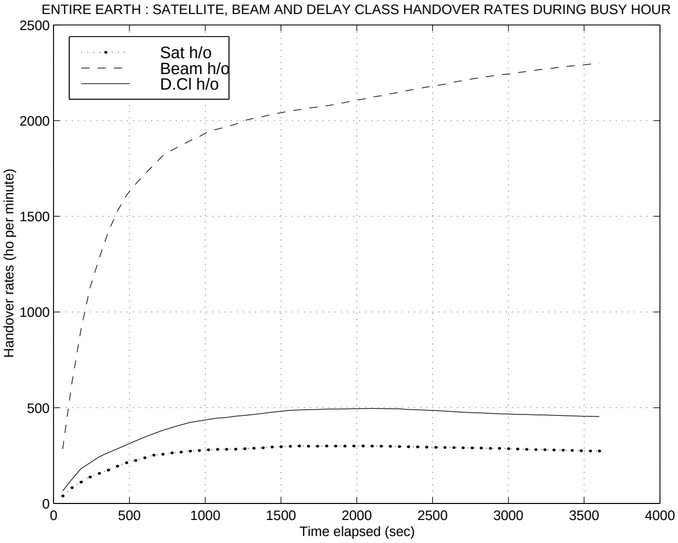

Figure 3.12: Aggregate satellite, beam and delay class handover rates on the earth.

Chapter 4

Joint base station and channel allocation policies in linear cellular networks

4.1 Introduction

4.1.1 Channel allocation policies

The limiting availability of radio spectrum imposes an inherent bound on the capacity of a mobile cellular system. To maximize system capacity, effort should focus on maximizing frequency reuse as much as possible. However, this may increase the mutual interferences among the cellular users. To maintain a certain quality of service, one has to keep the interference below a predefined level. This requirement is translated into some compatibility constraints. Allocating the channels in an efficient way which does not violate these constraints is the main objective of the Channel Assignment Problem (CAP).

Several methods for efficient Channel Allocation in cellular networks have been proposed and studied in the literature. In the simple formulation of the CAP, only cochannel interference is considered and the problem is known to be equivalent to the classical Graph Coloring problem. Owing to the NP-completeness of this problem [23], an exact search for the optimal solution is impractical for a large-scale system. Therefore, most of the efforts are spent in developping heuristics and approximation algorithms, a great portion of which are based on graph theoretic expansions. Cell topology, reuse constraints and cochannel interference limitations are depicted in graph modelling [24]- [28]. These algorithms can occasionally find optimal solutions, but in general provide only suboptimal solutions, with no information on how far away they are from the optimal one. Gamst [29] derives some lower bounds for the minimum number of channels required. An approach towards acquiring an analytical solution of blocking probabilities in linear cellular systems is presented in [30].

One of the most important Dynamic Channel Assignment (DCA) policies in cellular radio communication systems is the *Maximum Packing* policy, suggested by Everitt and Macfadyen [31]. The Maximum Packing (MP) policy specifies that a new call attempt is admitted, whenever there is some way of rearranging channels, so that every call can be supported. Although its practical use is limited because of a possible rearrangement of calls in progress on a global basis, it has several advantages. In cases where MP is analytically tractable, it can be used to compute exact blocking probabilities for a variety of system layouts and traffic patterns. The performance of MP in linear networks with a reuse cluster of two cells can be computed exactly by a recursive algorithm, as shown in [32]. Raymond [33] proved that in a linear network, it is possible to implement MP without ever doing more than two rearrangements upon arrival of a new call, independently of the size of the network. MP algorithm in that case entails minimum spectrum utilization. This proposition, however, cannot be extended to other reuse constraint or to two-dimensional general networks. For a general network of hexagonal cells, the worst case number of rearrangements between two subsequent arrivals in any implementation of MP grows without bound with the number of cells in the network.

Clique Packing (CP) was a policy also introduced by Raymond in [33], according to which a new call is accepted if and only if the mobiles' population in any of the cliques in the system does not exceed the total number of available channels in the system. A detailed treatment of CP is given in [34], where an approximation formula to calculate the blocking probability is derived.

4.1.2 Load Balancing

Load Balancing is a possible guiding principle for resource allocation, whereby the load is allocated across resources as evenly as possible. It should be noted however that in some situations it may not be efficient. For example, consider the situation of twelve users and two base stations, such that each base station can handle four or fewer users perfectly, but a base station fails completely if five or more users are assigned to it. Then it is preferable to allocate four users to one station and eight to the other (so that at least four users are satisfied) than to balance the load by assigning six users to each station, (so that no users are satisfied). It is known that Load Balancing can be an effective allocation strategy, when the associated cost is convex (or the reward concave) as a function of the allocated loads.

In [35] several routing policies for arriving customers towards a number of queues are proposed and compared. Those policies are periodic routing, Random routing, Join-the-Shortest-Queue (JSC) routing and also cyclic and time-varying routing.

In [36] the Static and Dynamic Load Balancing problems are defined. The dynamic allocation problem is formulated as an optimal control problem with a longterm average quadratic cost to be minimized within a set of practical controls. Least Loaded Routing (LLR) is a simple control, which assigns each call to the location with the lightest possible load. Asymptotic optimality of LLR is proved for large arrival rates. In the case of finite capacity resources, Least Relatively Loaded Routing (LRLR) is the regime which leads to minimization of blocking probability.

4.1.3 The joint base station and channel allocation problem

Overlapping coverage areas of nearby base stations arise naturally in cellular communication systems, especially in small-cell, high-capacity microcellular configurations. Calls arising in the common area of cells have access to channels at more than one base station. With an appropriate control strategy, a call may select the base station to establish a connection and contribute to efficient spectrum management. In this case the problems of base station and channel assignment arise jointly.

We address the problem in a linear cellular system, where cells are arranged in a linear array. Each base station provides coverage to a certain area, its cell, and the coverage areas of neighboring base stations are overlapping. A call generated in the overlap area of two base stations may choose which base station to use for the connection, while a call generated in an area from where a single base is only accessible will have to establish connection using that base.

We consider the base station and channel assignment problems jointly, aiming to a policy that will result in minimization of utilized number of channels. We present two algorithms in this context: The first one expands the Load Balancing principle in clique populations and will be hereafter referred to as Sequential Clique Load Balancing (SCLB). The second one can be visualized as an application of Inverse Water-Filling argument to clique population balancing and will be refered to as Clique Load Balancing with Inverse Water-Filling (CLB-IWF). These two algorithms are shown to be equivalent in effect. In a dynamic environment, we unify SCLB and CLB-IWF into CLB-DA, which comprises Dynamic Allocation. CLB-DA is compared with the classical Least Loaded Routing (LLR) and the Random Routing assignment policies. It can be concluded that CLB-DA outperforms classic LLR, attaining smaller blocking probability.

In section 4.2 we formulate the problem of minimum utilized spectrum as the minimization of the maximum clique and in section 4.3 we present basic definitions that will be used throughout our study. In section 4.4 we consider the problem in the context of Load Balancing and formulate it as a Quadratic Programming one. Section 4.5 describes the transition to SCLB and CLBIWF algorithms for static loads. Those algorithms are explicitly analyzed and compared in sections 4.6, 4.7 and 4.8. In section 4.9 the channel allocation procedure is investigated separately, as a phycical continuation of base station assignment. Section 4.10 extends the joint problem in a dynamic realistic environment. Finally, in section 4.11 comparative performance results between our method and some other methods are provided.

4.2 Problem formulation

When a call request appears in a region of a cellular network, the twofold question of its routing to the appropriate base station and its being assigned to a specific channel arises. First, the mobile should be assigned to the appropriate base station. Potential failure results most probably in call corruption due to insufficient signal strength. Accordingly, resources like timeslot and/or frequency have to be assigned to a call to establish connection.

We distinguish between *type-1* and *type-2* mobiles (calls). *Type-1* calls are definitely assigned to the Base Station in whose location area they belong and channels are assigned to them by some Dynamic or Fixed Channel Allocation technique (DCA or FCA). However, the situation is different with *type-2* calls which arise in an overlap area. In that case the mobile may be assigned to either of the base stations and at most a handover from one base station to the other will occur, without any further undesirable consequences to the call. Therefore, the determination of the base station to which a *type-2* mobile will be routed is itself pointless, unless a new goal to be fulfilled is introduced.

Base station allocation in that case can be realized in the context of efficient resource assignment. The number of utilized channels within a cell is directly affected by base station allocation of *type-2* mobiles and a least resource-consuming policy will result in the minimum number of channels being utilized. For a linear cellular network with reuse distance $R = 2$, channel reuse constraints instruct that a channel serving a call at a base station i cannot be simultaneously serving any call residing in any of the neighboring cells $i-1$ or $i+1$. The ultimate number of utilized channels in the system is equal to the maximum of the sums of the populations of any two adjacent cells. Therefore, the problem of allocating the mobiles to the appropriate base station with an outlook of achieving the minimum consumption of spectrum, can be mathematically stated as follows:

$$minimize \ \max_i (c_i + c_{i+1}) \tag{4.1}$$

where c_i and c_{i+1} are the populations of of two adjacent cells, which get enhanced upon assignment of a *type-1* or a *type-2* mobile to the corresponding base station.

4.3 Definitions

The following definitions are used throughout the chapter:

- *Linear Network* : A linear network is a cellular network where cells are arranged in a linear array (Fig 4.1).

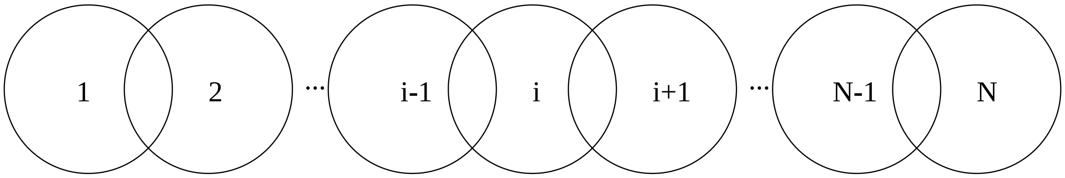

Figure 4.1: Cells arranged in a linear array.

- *Interference cells* : Consider cells u and v. They are said to interfere with each other if the use of a frequency in u prohibits the use of the same frequency in v and vice versa. In that case u is called an interference cell of v. Interference is obviously an equivalence relation between two cells.
- *Reuse Distance* : The reuse distance R between two cells i and j is the minimum cell separation index such that a channel can be simultaneously

used in i and j. In order for the latter to happen we should have $|i-j| \geq R$. It is assumed that cells $c \epsilon \mathcal{C}$ have been indexed.

- *Clique of cells* : A clique Q of cells is a subset of $\mathcal{C}$ with the property that each cell of Q is an interference cell of all cells in Q. For example, in a linear network with reuse distance $R = 2$ the clique i consists of cells i and $i + 1$.

- *Clique population* : We define the population q_i of clique i as the sum of the populations of the cells comprising the clique.

- *Maximum Clique of cells* : A Maximum Clique of cells is traditionally the clique containing the maximum number of cells. A maximum clique is always defined with reference to a reuse distance and a cell. In a linear network with fixed reuse distance $R = 2$ the maximum clique always consists of two cells. By convention we will define the maximum clique as that clique Q_i having the maximum population.

- *Types of mobiles* : The *type-1* mobiles are those for which there is no ambiguity as to which base station they will be assigned. *Type-2* mobiles on the contrary are situated on the overlap area between two cells.

- *Clique population vector* :

$$\underline{Q} = (q_1, q_2, \ldots q_{N-1}) \tag{4.2}$$

- The *cell population vector* :

$$\underline{C} = (c_1, c_2, \ldots c_N) \tag{4.3}$$

- The *overlap area population vector* :

$$\underline{\hat{C}} = (\hat{c}_1, \hat{c}_2, \ldots \hat{c}_{N-1}) \tag{4.4}$$

- The *remaining load vector* :

$$\underline{\ell} = (\ell_1, \ell_2, \ldots \ell_{N-1}) \tag{4.5}$$

4.4 Mathematical formulation of the problem

4.4.1 Notational conventions

We define the present problem in terms of a consumer demand network, in a way similar to that presented in [36], with some obvious modifications to comply with the clique environment in the network. The notational conventions are Any linear network can be visualized as a consumer demand network, denoted by $(\mathcal{U}, \mathcal{Q}, \mathcal{N})$. We define the following entities:

- N is the number of cells in the system, and $M = N - 1$ is the number of cliques.

- $\mathcal{U}$ will denote the set of *consumer types*, which, in our case is considered to be the set of *all disjoint regions* in the system. $\mathcal{U}$ is the union of two sets U_1 and U_2, which denote the non-overlap and overlap areas respectively.

- $\mathcal{Q}$ will denote the set of *locations*, which in our case is essentially the set of all *cliques* (cell pairs) i. Thus the i-th clique $i \epsilon \mathcal{Q}$ consists of the cells i and $i + 1$.

- A *demand* for this network is a vector $(m(u) : u\epsilon\mathcal{U})\ \epsilon\mathcal{R}_{+}^{\mathcal{U}}$ or $(\lambda(u) : u\epsilon\mathcal{U})\epsilon\mathcal{R}_{+}^{\mathcal{U}}$, where $R_{+}^{\mathcal{U}}$ is the set of nonnegative real valued vectors with the index set $\mathcal{U}$. The first notation is valid in the static case, while the second one is applicable in the dynamic case.

- An assignment $f_{v,u}$ meets the consumer demand m if :

$$\sum_{v} f_{v,u} = m(u)\ \forall u\epsilon\mathcal{U} \tag{4.6}$$

  where the summation is with respect to all cliques directly affected (loaded) by the assignment.

- The *base load* b or *a priori load* at each location $v\epsilon\mathcal{Q}$ is the already existing load at location v, before any assignment takes place. It denotes in essence the initial clique load $q_i = c_i + c_{i+1} + \hat{c}_i$

- The *load* x at each location $v\epsilon\mathcal{Q}$ corresponding to the assignment f is given by

$$x(v) = b(v) + \sum_{u\epsilon U} f_{v,u}\ ,\ v\epsilon Q \tag{4.7}$$

4.4.2 Proof of convergence

We formulate the Load Balancing problem when resources are only available in integral numbers of units, as it is applicable in the case of mobiles served by channels within the cells. The solution to this integral assignment problem is Let Φ_0 be a convex function on Z_{+}, the set of non-negative integers and define $J_0(f)$ for an assignment f by

$$J_0(f) = \sum_{v=1}^{M} \Phi_0(x(v)) \tag{4.8}$$

where x is the load vector. Consider now the problem of finding an assignment vector f that solves the following optimization problem:

Problem P_0:

minimize { $J_0(f)$: f is an integral assignment, meeting demand m }

Consider also the corresponding continuous-time problem:

Problem P:

minimize { $J_(f)$: f is an assignment, meeting demand m }

where $J(f)$ is the corresponding cost function for the continuous case. The sample variance of the load for a load vector x is

$$V(x) = \frac{1}{M}\left(\sum_{v \epsilon \mathcal{Q}} x^2(v)\right) - \overline{x}^2 \tag{4.9}$$

where M is the number of cliques in the system, $M = N - 1$.

$V(x)$ is minimized when f is a solution to problem P for $\Phi_0(k) = k^2$. Given the assignment f for problem (P), let $T_u f$ denote the new assignment that results by minimizing $J(f)$ with respect to $f_{v,u}$ for u and $(f_{v,u'} : u' \neq u, v \epsilon \mathcal{Q})$ fixed, subject to the constraint (4.6).

Let $f^{(0)}$ be an arbitrary assignment meeting demand m and let $(s_i)_{i \geq 1}$ be a sequence with $s_i \epsilon \mathcal{U}$ for all i. Define now a sequence of assignments, $\left(f^{(i)}\right)_{i \geq 0}$ recursively, by:

$$f^{(i+1)} = T_{s_i} f^{(i)} \tag{4.10}$$

Then the following Theorem from [37] holds for problem (P) :

Theorem 1 *The cost $J(f^{(i)})$ is monotone nonincreasing in i and converges to the minimum cost for Problem (P).*

The above is applied in the case of the SCLB algorithm, where a convergence to the optimum cost is achieved. The reduction of the integral assignment problem (P_0) to (P) from [37] is utilized.

4.4.3 Non-linear Programming formulation

The Clique Load Balancing principle can be shown to be equivalent to that of minimizing the variance of the clique loads (populations), given by equation (4.9)

Changing the notation from $x(v)$ to x_i, our Load Balancing problem can be formulated as follows:

$$\mathit{minimize}\ \frac{1}{M}\sum_{i=1}^{M} x_i^2 - \left(\frac{1}{M}\sum_{i=1}^{M} x_i\right)^2 \tag{4.11}$$

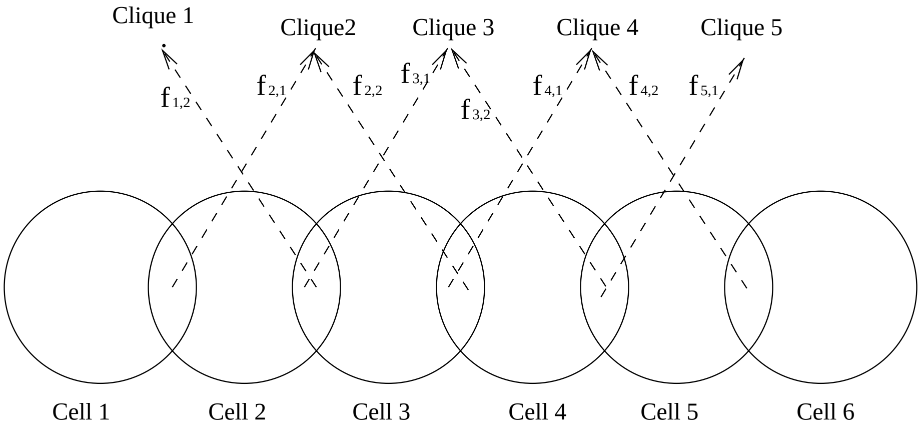

Figure 4.2: Clique assignment heuristic.

We consider the situation of the SCLB algorithm, where *only clique loading* is involved. Consider the situation depicted in figure 4.2. Each clique i, $i > 1$, can be loaded by the two adjacent overlap areas, namely the overlap area of mobiles $\hat{c}_{i-1}$ and that of mobiles $\hat{c}_{i+1}$. The first clique (indexed $i = 1$) can be loaded only from (a portion of) the overlap area of mobiles $\hat{c}_2$ and the last clique (indexed $i = N - 1$) is loaded only from (a portion of) the overlap area of mobiles $\hat{c}_{N-2}$.

For each clique i consider the assignment pair $(f_{i,1}, f_{i,2})$, where $f_{i,1}$ denotes the number of calls assigned to clique i from the adjacent overlap area *on the left* of the clique i and $f_{i,2}$ indicates the calls assigned to clique i from the adjacent overlap area *on the right* of the clique i (figure 4.2. Define now the $2 \times M$ assignment matrix:

$$F = \begin{pmatrix} f_{1,1} & f_{2,1} & \dots & f_{i,1} & \dots & f_{M,1} \\ f_{1,2} & f_{2,2} & \dots & f_{i,2} & \dots & f_{M,2} \end{pmatrix} \tag{4.12}$$

with $f_{1,1} = f_{N-1,2} = 0$, because cliques indexed 1 and $N-1$ form the boundaries of the system. Define also the *a priori load* vector $b(v), v\epsilon\mathcal{Q}$, which denotes the mobiles which unavoidably load that clique. For each clique i we can write the load x_i as

$$x_i = yFe_i + b_i = f_{i,1} + f_{i,2} + b_i \tag{4.13}$$

where $\underline{y}$ is the 1×2 vector $\underline{y} = [1 \ \ 1]$ and $\underline{e_i}$ is the unit coordinate vector of length M containing a "1" in the i-th position ans 0's elsewhere.

To pose the constraints for the problem, we recall the fact that a clique (location) is loaded only by the adjacent overlap areas. For such an overlap area with $\hat{c}_i$ mobiles, a portion $f_{i-1,2}$ is assigned to clique $i-1$ (base station i) and a portion $f_{i+1,1}$ is assigned to clique $i+1$ (base station $i+1$). We can therefore write

$$\hat{c}_i = f_{i-1,2} + f_{i+1,1} \ \ i = 2, 3, \dots, M-1 \tag{4.14}$$

$$\hat{c}_1 = \alpha + f_{2,1} \ \ , \ \ \hat{c}_{N-1} = f_{N-2,2} + \beta \tag{4.15}$$

where α and β are slack variables. The physical meaning of α is that not all of $\hat{c}_1$ is assigned to clique 2, since some calls are assigned to base station 1, which belongs to clique 1 and similarly can the meaning of the variable β be derived.

Using the above equations we can write the variance as follows:

$$\begin{aligned} V\left(\underline{f_1},\underline{f_2}\right) = & \quad \tfrac{1}{M}\left(\sum_{i=1}^{M} b_i^2 + \sum_{i=1}^{M}\left(f_{i,1}+f_{i,2}\right)^2 + 2\sum_{i=1}^{M} b_i\left(f_{i,1}+f_{i,2}\right)\right) \\ -\frac{1}{M^2} & \left[\left(\sum_{i=1}^{M} b_i\right)^2 + \left(\sum_{i=1}^{M}\left(f_{i,1}+f_{i,2}\right)\right)^2 + 2\sum_{i=1}^{M} b_i \sum_{i=1}^{M}\left(f_{i,1}+f_{i,2}\right)\right] \end{aligned} \tag{4.16}$$

Our goal is to formulate the above equation in terms of the $M \times 1$ vectors b, x so that an expression suited for Mathematic Programming can be obtained. Fix attention to term $\Delta_{\underline{b},\underline{b}} = \left(\sum_{i=1}^{M} b_i\right)^2$. Expand the above term and arrange terms as follows:

$$\begin{aligned} \Delta_{\underline{b},\underline{b}} = \; & b_1^2 + b_2^2 + \ldots + b_{M-1}^2 + b_M^2 + \\ & + b_2 b_1 + b_3 b_2 + \ldots + b_M b_{M-1} + b_1 b_M + \\ & + b_3 b_1 + b_4 b_2 + \ldots + b_M b_{M-2} + b_1 b_{M-1} + b_2 b_M + \\ & + \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots + \\ & + b_{M-1} b_1 + b_M b_2 + b_1 b_3 + b_2 b_4 + \ldots b_{M-3} b_{M-1} + b_{M-2} b_M + \\ & + b_M b_1 + b_1 b_2 + b_2 b_3 + \ldots + b_{M-2} b_{M-1} + b_{M-1} b_M + \end{aligned} \tag{4.17}$$

or

$$\Delta_{\underline{b},\underline{b}} = \underline{b}^T\underline{b} + \underline{b}^{1^T}\underline{b} + \underline{b}^{2^T}\underline{b} + \ldots + \underline{b}^{(M-1)^T}\underline{b} \tag{4.18}$$

where

$$\underline{b}^i = \left(b_{i+1}, b_{i+2}, \ldots, b_M, b_1, \ldots, b_i\right)^T \quad , \quad \underline{b}^0 = \underline{b} \tag{4.19}$$

Next observe that the “shifted” vectors $\underline{b}^i$ are produced by the basic vector $\underline{b}$ with the following rule :

$$\underline{b}^i = E_i \underline{b} \tag{4.20}$$

where

$$E_i = \begin{pmatrix} O_{(M-i)\times i} & I_{(M-i)\times(M-i)} \\ I_{i\times i} & O_{i\times(M-i)} \end{pmatrix} \tag{4.21}$$

Notice also that

$$\sum_{i=0}^{M-1} E_i = \Lambda \tag{4.22}$$

where Λ is a $M \times M$ symmetric matrix with 1's everywhere. We deduce therefore that

$$\Delta_{\underline{b},\underline{b}} = \underline{b}^T \Lambda \underline{b} \tag{4.23}$$

Similarly, by defining and calculating the entities

$$\Delta_{\underline{f}_k,\underline{b}} = \left(\sum_{i=1}^{M} f_{i,k}\right)\left(\sum_{i=1}^{M} b_i\right) = \underline{f}_k^T \Lambda \underline{b}, \quad k = 1, 2 \tag{4.24}$$

and

$$\Delta_{\underline{f}_k,\underline{f}_\ell} = \left(\sum_{i=1}^{M} f_{i,k}\right)\left(\sum_{i=1}^{M} f_{i,\ell}\right) = \underline{f}_k^T \Lambda \underline{f}_\ell \;\; k, \ell = 1, 2 \tag{4.25}$$

we can express the variance $V(x)$ in the following way:

$$\begin{aligned} V\left(\underline{f_1}, \underline{f_2}\right) = & \quad \tfrac{1}{M}\left(\underline{b}^T \underline{b} + \underline{f_1}^T \underline{f_1} + \underline{f_2}^T \underline{f_2} + 2\underline{f_1}^T \underline{f_2} + 2\underline{b}^T \underline{f_1} + 2\underline{b}^T \underline{f_2}\right) \\ -\frac{1}{M^2} & \left(\underline{b}^T \Lambda \underline{b} + \underline{f_1}^T \Lambda \underline{f_1} + \underline{f_2}^T \Lambda \underline{f_2} + 2\underline{f_1}^T \Lambda \underline{f_2} + 2\underline{b}^T \Lambda \underline{f_1} + 2\underline{b}^T \Lambda \underline{f_2}\right) \end{aligned} \tag{4.26}$$

Now set $\underline{f_1} = \underline{x}$ and $\underline{f_2} = \underline{y}$. Then the initial problem of minimizing the clique load variance V takes the following form:

$$\begin{aligned} & minimize \qquad V\left(\underline{x}, \underline{y}\right) \\ & subject \;\; to: \;\; A\underline{x} + B\underline{y} = \hat{C} \end{aligned} \tag{4.27}$$

where A, B are appropriately defined, and

$$\begin{aligned} V\left(\underline{x}, \underline{y}\right) = & \quad \tfrac{1}{M}\left(\underline{b}^T \underline{b} + \underline{x}^T \underline{x} + \underline{y}^T \underline{y} + 2\underline{x}^T \underline{y} + 2\underline{b}^T \underline{x} + 2\underline{b}^T \underline{y}\right) \\ -\frac{1}{M^2} & \left(\underline{b}^T \Lambda \underline{b} + \underline{x}^T \Lambda \underline{x} + \underline{y}^T \Lambda \underline{y} + 2\underline{x}^T \Lambda \underline{y} + 2\underline{b}^T \Lambda \underline{x} + 2\underline{b}^T \Lambda \underline{y}\right) \end{aligned} \tag{4.28}$$

This problem is identified as a Quadratic Programming problem with a linear constraint. Solving the equality constraint in terms of $\underline{y}$ we get

$$\underline{y} = B^{-1}\left(\hat{C} - A\underline{x}\right) \tag{4.29}$$

The existence of the inverse matrices can be disputed by the following fact: For cliques 1 and $N-1$ there are no contributions from the overlap regions located on the left and on the right of them, namely, $f_{1,1} = f_{N-1,2} = 0$, and therefore A and/or B have a row of zeros, a fact that excludes the existence of A^{-1} or B^{-1}. We circumvent this restriction by assuming that inverses exist, as there will exist some portion of mobiles α and β that will be assigned to those base stations. The above problem can then obtain the form of a Quadratic Programming (QP) problem

$$\begin{aligned} &minimize \quad V(\underline{x}) \\ &subject\ to \quad \underline{x} \geq \underline{0} \end{aligned} \tag{4.30}$$

with $V(\underline{x})$ of the form

$$V(\underline{x}) = \underline{x}^T Q \underline{x} + P\underline{x} + R \tag{4.31}$$

where

$$\begin{aligned} Q = & \quad \tfrac{1}{M}\left[I + A^T\left(B^{-1}\right)^T B^{-1}A\right] \\ - & \quad \tfrac{1}{M^2}\left[\Lambda + A^T\left(B^{-1}\right)^T \Lambda B^{-1}A - 2\Lambda B^{-1}A\right] \end{aligned} \tag{4.32}$$

$$\begin{aligned} P = & \tfrac{1}{M}\left[-\left(\hat{C}\right)^T\left(B^{-1}\right)^T A + 2\underline{b}^T - 2\underline{b}^T B^{-1}A - \left(\hat{C}\right)^T\left(B^{-1}\right)^T B^{-1}A\right] \\ & -\tfrac{1}{M^2}\left[2\underline{b}^T\Lambda - 2\underline{b}\Lambda B^{-1}A + 2\left(\hat{C}\right)^T\left(B^{-1}\right)^T\Lambda\right] \end{aligned} \tag{4.33}$$

$$\begin{aligned} R = & \quad \tfrac{1}{M}\left[\underline{b}^T\underline{b} + 2\underline{b}^T B^{-1}\hat{C} + \left(\hat{C}\right)^T\left(B^{-1}\right)^T B^{-1}\hat{C}\right] \\ - & \quad \tfrac{1}{M^2}\left[\underline{b}^T\Lambda\underline{b} + \left(\hat{C}\right)^T\left(B^{-1}\right)^T\Lambda B^{-1}\hat{C} + 2\underline{b}^T\Lambda B^{-1}\hat{C}\right] \end{aligned} \tag{4.34}$$

The QP problem can be solved with some standard Non-Linear Programming or Quadratic Programming algorithm, such as Convex-Simplex Method, Conjugate Direction (CSM-CD) or the Complementary Pivoting algorithm [38], [39]. The continuous, non-integer solution can then be rounded to the closest integer to comply with the integer assignment constraint problem.

4.5 The transition to SCLB and CLBIWF algorithms

In the previous section we formulated the Clique Load Balancing problem as a Quadratic Programming one, which can be solved by standard methods. One could argue that in a realistic environment, this problem can be encountered in the same way, by solving succesive Quadratic Programming problems at every time instant. This approach however is impractical for a system with a relatively high number of cells, as complexity of computations rises exponentially. An algorithm suited to work in a real-time environment is required. To this end, the above mentioned SCLB and CLBIWF algorithms come into stage.

Consider a linear network consisting of N cells, indexed with integer numbers 1 to N. Assume also that a population of mobiles $c_1, c_2 \ldots c_N$ and $\hat{c}_1, \hat{c}_2 \ldots \hat{c}_{N-1}$ is established. Concerning the definition of the *clique population* for the clique Q_i , two alternatives exist:

$$q_i = c_i + c_{i+1} + \hat{c}_i \tag{4.35}$$

$$q_i = c_i + c_{i+1} + \hat{c}_{i-1} + \hat{c}_i + \hat{c}_{i+1} \tag{4.36}$$

The above duality arises from the question of base station assignment of *type-2* mobiles. The key fact is that those mobiles cannot be counted as mobiles belonging to a clique *unless they are ultimately assigned to one of the two base stations.* For the clique i it is evident that $\hat{c}_i$ mobiles should be definitely counted towards the clique population, since those mobiles will be assigned to either of the base stations i or $i+1$. This is not the case however for mobiles $\hat{c}_{i-1}$ and $\hat{c}_{i+1}$. Mobiles $\hat{c}_{i-1}$ may be assigned either to base $i-1$ (thus affecting clique $i-2$) or to base i (thus affecting clique i). Similarly, mobiles $\hat{c}_{i+1}$ are assigned either to base $i+2$ (thus affecting clique $i+2$) or to base $i+1$ (thus affecting clique i).

SCLB algorithm considers the first alternative for clique populations, while CLBIWF assumes the second alternative to hold. The performance of both algorithms is analyzed in the following sections.

For each clique i and before the k-th algorithm iteration we define the following quantities:

- The clique population $q_i^{(k)}$
- The base station population $c_i^{(k)}$, with $c_i^{(0)} = c_i$
- The overlap region population $\hat{c}_i^{(k)}$, with $\hat{c}_i^{(0)} = \hat{c}_i$
- The load that needs to be removed from a clique, $\ell_i^{(k)}$, so that no mobile remains unassigned.
- The cost function

$$J^{(k)} = \sum_i \left(q_i^{(k)}\right)^2 \tag{4.37}$$

4.6 The SCLB Algorithm for static loads

Taken into consideration the existing channel reuse constraints in the system, the number of channels that will finally be utilized is determined from the maximum clique population. To minimize the number of channels, an efficient policy should focus on appropriate assignment of *type-2* mobiles, since only the assignment of those mobiles influences clique populations.

We propose an algorithm which is based on the idea of Load Balancing. The clique is assumed to have a predefined number of mobiles: the mobiles assigned to each of the two base stations of the clique and the mobiles generated in the corresponding overlap region, according to equation 4.35. In addition to those, a number of mobiles, generated by a splitting mechanism, augments the clique population each time. Those mobiles are generated in the overlap areas adjacent to the clique, according to figure 4.2 The splitting mechanism decides how many of those will be assigned to each clique.

We begin from a random initial splitting of the mobiles to the base stations, so that each base station and each clique are assigned an initial population. It can be proved that this initial assignment does not affect the ultimate convergence of the algorithm. At each iteration the assigned mobiles are subtracted from the corresponding cells and they are reassigned to them, so that cliques are locally balanced. The algorithm terminates after a finite number of iterations, when all clique populations remain unaltered after two subsequent assignments.

4.6.1 Algorithm Description

For each iteration k and for each clique i, define a number of mobiles $y_i^{(k),+}$, denoting the number of mobiles assigned to clique i (cell $i+1$) from its right overlap area of this clique (i.e from the area containing $\hat{c}_{i+1}$ calls) and a number of mobiles $y_i^{(k),-}$, denoting the number of mobiles assigned to clique i (cell i) from its left overlap area (i.e from the area containing $\hat{c}_{i-1}$ calls). The following equality should obviously be satisfied at all iterations k

$$y_{i-1}^{(k),+} + y_{i+1}^{(k),-} = \hat{c}_i \tag{4.38}$$

and the clique population at iteration k will be:

$$q_i^{(k)} = c_i^{(k)} + c_{i+1}^{(k)} + \hat{c}_i + y_i^{(k),+} + y_i^{(k),-} \tag{4.39}$$

The consecutive steps of the algorithm are presented below:

1. In the original setup we assign the initial clique populations as follows:

$$q_i^{(0)} = c_i + c_{i+1} + \hat{c}_i \quad i = 1, 2, 3 \dots N-1 \tag{4.40}$$

2. For each overlap region population $\hat{c}_i$, produce a random number $y_i < \hat{c}_i$ and assign y_i mobiles to base station i (clique $i-1$) and the rest $\hat{c}_i - y_i$ to base $i+1$ (clique $i+1$). The populations of the affected cells and cliques are updated accordingly. The overlap areas indexed 1 and $N-1$ require a special treatment:

   - For the overlap area indexed 1, produce a random number $y_1 < \hat{c}_1$ and assign y_1 mobiles to base station 1 (clique 1) and the rest $\hat{c}_1 - y_1$ to base 2 (clique 2), noting that in that case only clique 2 receives additional load.

- For the overlap area indexed $N-1$, produce a random number $y_{N-1} < \hat{c}_{N-1}$ and assign y_{N-1} mobiles to base station $N-2$ (clique $N-2$) and the rest $\hat{c}_{N-1} - y_{N-1}$ to base $N-1$ (clique $N-1$), noting that in that case only clique $N-2$ receives additional load.

3. From all cells remove the assigned mobiles from the previous iteration and therefore reduce the clique populations:

$$c_i \rightarrow c_i - y_{i-1}^{(k-1),+} \tag{4.41}$$

$$q_{i-1} \rightarrow q_{i-1} - y_{i-1}^{(k-1),+} \tag{4.42}$$

$$c_{i+1} \rightarrow c_{i+1} - y_{i+1}^{(k-1),-} \tag{4.43}$$

$$q_{i+1} \rightarrow q_{i+1} - y_{i+1}^{(k-1),-} \tag{4.44}$$

4. For each overlap area containing mobiles $\hat{c}_j$, perform steps 5-7.

5. Choose the minimum of the adjacent cliques, $min\{q_{j-1}, q_{j+1}\}$. Suppose for example that $q_{j-1} < q_{j+1}$.

6. Start assigning mobiles to clique $j-1$ (base station j) until the two clique populations are equal to each other, or until all mobiles $\hat{c}_j$ are exhausted.

7. If after executing step 6, there are mobiles left in the overlap area, start assigning those interchangeably to cliques q_{j-1} and q_{j+1}.

8. Record the clique population vector $Q^{(k)}$ after the end of the k-th iteration and compare it to the corresponding vector $Q^{(k-1)}$ after the previus iteration.

9. Terminate the algorithm if $\underline{Q}^{(k)} = \underline{Q}^{(k-1)}$.

4.7 The CLBIWF Algorithm for static loads

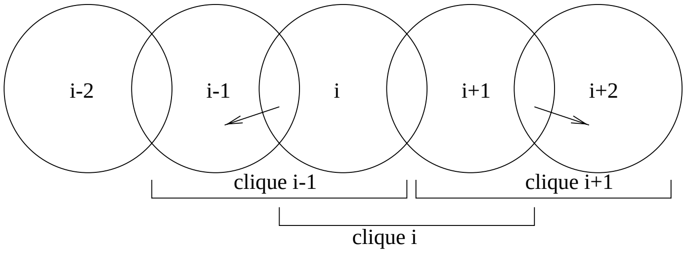

Figure 4.3: Clique unloading

The CLBIWF algorithm is based on the idea of Clique Load Balancing through clique unloading. A call among $\hat{c}_i$ belongs *to both cliques* $i-1$ and $i+1$. At each iteration of the algorithm, the clique with the maximum population is detected (say clique i). Accordingly, clique i gets unloaded by assigning the overlap area mobiles to adjacent base stations *not included* in the present clique, i.e base stations $i-1$ and $i+2$, in our case. The clique is unloaded until the next higher clique load level is reached and the algorithm proceeds similarly in each iteration, by unloading the clique(s) with the higher population. The procedure ends when all cliques have been unloaded and there exists no unassigned mobile left. The procedure obviously resembles Inverse Water-filling, a term established from Information Theory [40].

Clique unloading itself is feasible by selecting the greater of the two adjacent overlap areas, in terms of the remaining load they contain, and accordingly assign them to the base stations not included in the clique. This policy ensures the even and smooth clique unloading. If one of the two adjacent overlap areas contains no load to be removed, only the other overlap area is unloaded. Each time an unloading event takes place, all changes involving increase in certain cell populations or decrease in clique loads are recorded. Of particular importance is the fact that *no clique is loaded*, since all mobiles in overlap areas belong to all cliques $i-1$, i and $i+1$. After termination of the algorithm, the system is "balanced" as far as clique populations is concerned.

4.7.1 Algorithm description

In the original setup we assign the initial clique populations and loads as follows:

$$q_i^{(0)} = c_i + c_{i+1} + \hat{c}_{i-1} + \hat{c}_i + \hat{c}_{i+1},\ i = 2, 3 \dots N-2 \tag{4.45}$$

$$q_1^{(0)} = c_1 + c_2 + \hat{c}_1 + \hat{c}_2 \tag{4.46}$$

$$q_{(N-1)}^0 = c_{N-1} + c_N + \hat{c}_{N-2} + \hat{c}_{N-1} \tag{4.47}$$

while for the loads we have the initial conditions

$$\ell_i^{(0)} = \hat{c}_{i-1} + \hat{c}_{i+1}\ \ i = 2, 3 \dots N-2 \tag{4.48}$$

$$\ell_1^{(0)} = \hat{c}_1 \tag{4.49}$$

$$\ell_{N-1}^{(0)} = \hat{c}_{N-2} \tag{4.50}$$

The algorithm consists of the following steps:

1. For each algorithm iteration k do steps 2-5.

2. Sort the clique populations of all cliques i for which there exists load to be removed, i.e $\ell_i^{(k)} > 0$.

3. Obtain an ordering of those cliques, $q_{(1)}, q_{(2)}, \dots q_{(m_k)}$ in non-decreasing order, where m_k is the number of cliques such that $\ell_i^{(k)} > 0$.

4. Find the maximum clique index (indices)

$$i^* = \max_{i:\ell_i^{(k)}>0} q_i^{(k)} \tag{4.51}$$

   where $q_{i^*} = q_{(1)}$.

5. Start unloading this clique, until the next higher population level is reached, or until all the load is exhausted, i.e until $q_{(1)} = q_{(2)}$, or $\ell_{i^*}^{(k)} = 0$. Unloading is achieved according to the procedure CLIQUE UNLOADING.

6. Continue with iteration $k+1$ and stop at iteration k^* when $\ell_i^{(k^*)} = 0$ for all i.

The CLIQUE UNLOADING procedure at the k-th iteration consists of the following steps (assume $i = i^*$):

To decide on which of the two overlap areas corresponding to the clique to unload, find the maximum of the populations of these areas, i.e find $\max\{\hat{c}_{i-1}, \hat{c}_{i+1}\}$. There exist two cases:

- Case I: $\hat{c}_{i-1} > 0$, $\hat{c}_{i+1} > 0$ and $\hat{c}_{i-1} > \hat{c}_{i+1}$ OR $\hat{c}_{i+1} = 0$

- Case II: $\hat{c}_{i-1} > 0$. $\hat{c}_{i+1} > 0$ and $\hat{c}_{i-1} < \hat{c}_{i+1}$, OR $\hat{c}_{i-1} = 0$,

For case I we continue as follows:

1. Assign a mobile from $\hat{c}_{i-1}$ to base station $i-1$, with the following impact:

2. Reduce this overlap area population by 1,

$$\hat{c}_{i-1}^{(k+1)} = \hat{c}_{i-1}^{(k)} - 1 \tag{4.52}$$

3. Reduce the clique population by 1,

$$q_i^{(k+1)} = q_i^{(k)} - 1 \tag{4.53}$$

4. Reduce the load that clique to be removed by 1,

$$\ell_i^{(k+1)} = \ell_i^{(k)} - 1 \tag{4.54}$$

5. Increase the number of mobiles in the base station where the call was assigned:

$$c_{i-1}^{(k+1)} = c_{i-1}^{(k)} + 1 \tag{4.55}$$

6. If $i > 2$ then denote the transition by reducing also the load of the other affected clique,

$$\ell_{i-2}^{(k+1)} = \ell_{i-2}^{(k)} - 1 \tag{4.56}$$

For case II we proceed in a similar fashion with case I

1. Assign a mobile from $\hat{c}_{i+1}$ to base station $i+2$, with the following impacts:

2. Reduce this overlap area population by 1,

$$\hat{c}_{i+1}^{(k+1)} = \hat{c}_{i+1}^{(k)} - 1 \tag{4.57}$$

3. Reduce the clique population by 1,

$$q_i^{(k+1)} = q_i^{(k)} - 1 \tag{4.58}$$

4. Reduce the load of that clique to be removed by 1,

$$\ell_i^{(k+1)} = q_i^{(k)} - 1 \tag{4.59}$$

5. Increase the number of mobiles in the base station where the call was assigned:

$$c_{i+2}^{(k+1)} = c_{i+2}^{(k)} + 1 \tag{4.60}$$

6. If $i < N-2$ then denote the transition by reducing also the load of the other affected clique,

$$\ell_{i+2}^{(k+1)} = \ell_{i+2}^{(k)} - 1 \tag{4.61}$$

In the case of the boundary clique 1, only the mobiles $\hat{c}_2$ constitute the load to be removed from the clique and similarly for clique $N-1$, only mobiles $\hat{c}_{N-2}$ can be removed. Those special cases should be taken into consideration.

4.8 Comparison of SCLB and CLBIWF

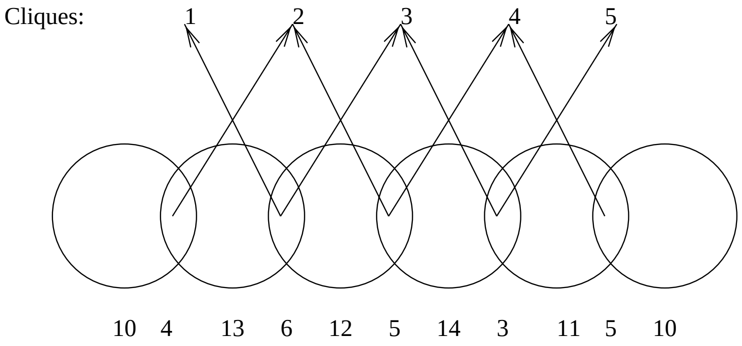

Figure 4.4: The linear network for the comparison of the two algorithms

We consider a linear cellular network consisting on $N = 6$ cells and cell and overlap area populations as shown in figure 4.4. The results obtained after the execution of SCLB and CLBIWF algorithms are shown in tables 4.1 and 4.2 at the end of the chapter. It turns out that the final clique population vector is identical for both cases, $\underline{Q} = (32, 32, 32, 32, 29)$. Therefore, the minimum number of channels that should be utilized are 32 for both cases. In general, the performance of the two algorithms is the same with respect to the minimum number of channels.

The two algorithms are based on different initializations but converge to the same result. SCLB algorithm assumes that the effect of the overlap region assignments to a base station is the *loading* of two cliques and no clique is unloaded. On the other hand, CLBIWF algorithm implements the overlap region assignments by *unloading* two cliques (or one, in the boundaries). The SCLB algorithm itself is however of lower complexity and also appears more flexible in being applied in the case of a real-time environment.

Let us focus on the behavior of the cost function $J^{(k)}$. In the case of the CLBIWF algorithm, it is obvious that $J^{(k)}$ is non-increasing with k, as cliques get unloaded, and there the convergence is evident. For the SCLB algorithm in the example of figure 4.4, we recorded the following evolution of cost, which serves as a verification to theorem 1.

$$\begin{array}{ccccccccc} 5267 & \rightarrow & 5227 & \rightarrow & 5209 & \rightarrow & 5203 & \rightarrow & 4947 \\ 4947 & \rightarrow & 4945 & \rightarrow & 4937 & \rightarrow & 4937 & END & \end{array}$$

4.9 Channel allocation

Assume that the first half of our objective has been realized, namely all mobiles have been assigned to a base station in the above analyzed context of Clique Load Balancing. At that stage, clique populations can be determined with accuracy and are balanced in the maximum possible extent. The next task is to allocate channels to mobiles, taking into consideration the base station they belong to. We will show that the number of channels required is exactly

$$N_{CH} = \max_{i=1\ldots N-1} q_i \tag{4.62}$$

where

$$q_i = c_i + c_{i+1} \tag{4.63}$$

because now there are no unassigned mobiles in the overlap areas. We claim that any mobile population configuration in the cells can be accomodated by using this number of channels.

Assume that $i = i^*$, so that $q_i^* = \max_i q_i$. Divide the set of indexed cells into two disjoint subsets

$$\begin{aligned} C_1 &= c_{i^*-2\kappa} \ldots, c_{i^*-2} c_{i^*}, c_{i^*+2} \ldots c_{i^*+2\kappa} \\ C_2 &= c_{i^*-2\kappa+1} \ldots, c_{i^*-1}, c_{i^*+1} \ldots c_{i^*+2\kappa+1} \end{aligned} \tag{4.64}$$

Given the reuse constraint $R = 2$, channels may be reused within the set C_1 or within the set C_2 of cells. It is straightforward that, after completion of the aforementioned base station allocation strategy, there are two possible cases with regard to mobile population configuration.

- CASE A:

$$c_{i^*} = \max_{\kappa} \{c_{i^*\pm 2\kappa}\} \ AND \ c_{i^*+1} = \max_{\kappa} \{c_{i^*\pm 2\kappa+1}\} \tag{4.65}$$

In this case, a portion of the c_{i^*} channels will be reused in cells $c_{i^*\pm 2\kappa}$ and a portion of the c_{i^*+1} channels will be reused to cells $c_{i^*\pm 2\kappa+1}$. For example, consider a linear network of four cells after the base station allocation, so that

$$c_1 = 15, \ c_2 = 16, \ c_3 = 17, \ c_4 = 18 \tag{4.66}$$

The number of channels utilized is determined by the maximum clique q_3 and is 35. For calls in cell 1, 15 out of 17 channels will be reused, whereas for calls in cell 2, 16 out of 18 channels are reused.

- CASE B:

$$c_{i^*} = \max_{\kappa} \{c_{i^*\pm 2\kappa}\} \ OR \ c_{i^*+1} = \max_{\kappa} \{c_{i^*\pm 2\kappa+1}\}, \ but \ NOT \ both \tag{4.67}$$

Notice that at least one of these conditions hold, otherwise the maximum clique would be other than q_{i^*}. Assume now that

$$c_{i^*} < \max_{\kappa} \{c_{i^* \pm 2\kappa}\} \tag{4.68}$$

In that case, the allocation procedure is slighly more complicated. The standard reuse procedure is followed for cells $c_{i^*+2\kappa+1}$. The assumption that *only one* of the above conditions holds, but *not both*, is equivalent to saying that there exists at least one cell index m, even, such that $c_m > c_{i^*}$. The fact that q_{i^*} is the maximum clique means that $c_{m-1} < c_{i^*+1}$ and $c_{m+1} < c_{i^*+1}$, otherwise the maximum clique would be q_{m-1} or q_m respectively. Since

$$c_{i^*+1} - c_{m+1} \geq c_m - c_{i^*} \tag{4.69}$$

$$c_{i^*+1} - c_{m-1} \geq c_m - c_{i^*} \tag{4.70}$$

a portion of the unreused channels $c_{i^*+1} - c_{m+1}$ (or the $c_{i^*+1} - c_{m-1}$) can be used for the $c_m - c_{i^*}$ calls of cell m. The rest of the mobiles in cell m can be assigned reused channels.

For example for a linear network with five cells, such that: $c_1 = 17$, $c_2 = 18$, $c_3 = 16$, $c_4 = 17$ and $c_5 = 18$, we see that the maximum clique is q_4 with 35 channels. We reuse 16 out of the 18 channels for calls of cell 3 and 17 out of 18 channels for calls of cell 1. Now, for the 18 calls of cell 2, reuse the 17 channels of cell 4. For the remaining one mobile, use the channel belonging to the set of 18 channels of cell 5 that has not been used neither to cell 3 nor to cell 1.

4.10 Algorithms for joint Base Station and channel allocation for dynamic loads

In the previous sections the joint problem of base station and channel allocation has been considered in a static environment, namely a fixed number of mobiles was considered for each region in the linear cellular network and the SCLB and CLBIWF algorithms resulted in the accomodation of all mobiles in the base stations, so that the minimum number of channels was utilized. It was deduced that the performance of the two algorithms is equivalent.

In a real environment however, everything evolves dynamically. Call arrivals and departures from the system obey a specific policy and the problem now becomes to assign *type-2* mobiles to base stations, taking into consideration their occupancy in a specific time instant. The goal remains the minimization of the consumed resources in the network, however, the issue of call blocking probability comes into question. Ideally, we would like to optimize network performance with respect to both the aforementioned parameters, but this is not feasible, as extra bandwidth must be sacrificed in order to ensure higher level Quality of Service (which translates into low enough blocking probability).

SCLB and CLBIWF algorithms are merged into one algorithm, the Clique Load Balancing-Dynamic Allocation (CLB-DA) algorithm. This unification is absolutely legitimate, since the two algorithms are shown to be equivalent in effect. Least Loaded Routing (LLR) allocation policy performance for dynamic loads is also depicted and compared to CLB-DA algorithm. It can be deduced that CLB-DA outperforms LLR. The other simple algorithm applied for dynamic loads involves static allocation and will be hereafter referred to as Static

Allocation (SA) or Random routing algorithm.

The extension of the algorithms to the dynamic case is obvious. Dynamic Allocation procedure involves sequential assignment and reassignment of calls of the overlap regions to the neighboring base stations, using Clique Load Balancing as a guiding rule. This procedure takes place at any time instant t_k and considers the system parameters and traffic at that specific time point.

LLR Allocation procedure is similar to the one proposed in [Hajek]. A mobile in the overlap region is routed *only once* to this base station which currently serves fewer calls. The selection of the appropriate base station is made out of the two neighboring base stations which are eligible to serve the call.

Static Allocation simply assigns a call of the overlap region in one of the neighboring base stations, which are eligible to offer support to the call. The assignment is random and is made *only once*.

4.10.1 Call blocking

After the basic setup, we consider the blocking of a call, which is caused owing to finite capacity of the system. Calls are generated independently according to a Poisson distribution with the arrival rates appropriately scaled to account for the overlap and the non-overlap region traffic intensities and they are attached to a list, in which all active calls on the system are maintained. To investigate the blocking probability of a call, we need to distinguish between the case when a call is generated in an overlap or non-overlap region and whether it is located in a middle or an edge area (overlap or non-overlap). It becomes obvious that blocking is related to clique population at the time the call arrives in the system. Therefore, we define the event K_i as the event that the population of the i-th clique exceeds N_{CH}, i.e

$$K_i = \{q_i \geq N_{CH}\} \tag{4.71}$$

Consider a call generated in a non-overlap region (cell) i such that $1 < i < N$, i.e the call is not located in either of the two boundary regions. Upon arrival, this call will unavoidably load cliques $i-1$ and i. Therefore, it is blocked and removed from the system when either of the two cliques is full, or when both are full. In other words, there exists blocking whenever cone (or both) cliques $i-1$, i have the maximum population, namely when the event

$$\Omega_1 = K_i \cup K_{i-1} \tag{4.72}$$

takes place, where N_{CH} is the maximum number of channels in the system and is reflected on the population of the maximum clique.

In the case where the call is generated in the first (last) non-overlap regions, blocking occurance is equivalent to declaring that *only* event K_1 (*only* event K_{N-1}) suffices to consider a call blocked.

Calls generated in overlap areas are considerably fewer than those in non-overlap areas but they are treated differently, due to the fact that there exists ambiguity as to which clique will be loaded. Only after the assignment algorithm is executed, is the loaded cell (and consequently, the loaded clique) determined. Assume again, as before that the overlap region j is such that $1 < j < N-1$. Then the call may be assigned to cell j (in which case clique $j-1$ will be loaded) or to cell $j+1$ (in which case clique $j+1$ will be loaded). Apparently, clique j will always be affected. If $q_j < N_{CH}$, it is proved that there always exists a channel arrangement so that the call will be accomodated within the specified bandwidth. As a result, the call will be blocked if there exists no channel available to be allocated to it. But even if clique j is not the maximum clique, the call may be blocked if both the adjacent cliques $j-1$ and $j+1$ are full. Therefore, in that case, blocking occurs when the following event occurs:

$$\Omega_2 = K_j \cup (K_{j-1} \cap K_{j+1}) \tag{4.73}$$

Regarding calls in the first (last) overlap region, *only* events K_1 (K_{N-1}) can register a call as blocked. For the first overlap region, if $q_1 < N_{CH}$ and $q_2 \geq N_{CH}$ then, by construction, the call will be routed to cell 1 and clique 2 will not be affected. A similar situation holds for the last overlap region.

Upon arrival of a new call in the system, potential call completions in all cells are recorded and the cell and clique populations are updated. Accordingly, the main algorithm takes place, whereby calls are ultimately routed to the appropriate cells in such a way that the minimum bandwidth is consumed. Only the recently arrived unassigned calls in the overlap regions must be considered for assignment, while ongoing calls in overlap regions are in effect counted towards clique populations, according to the base stations they are assigned to. The blocking probability is simply defined as

$$P_B = \frac{N_B}{N_m} \tag{4.74}$$

where N_B is the number of mobiles that experience blocking and N_m is the total number of mobiles trying to enter the system and establish a connection.

4.11 Results and extensions

To backup the analysis of the previous sections, a real-time simulation environment was built, where the aforementioned allocation algorithms were compared as far as their performance is concerned. The simulation environment was coded in C.

We assume that traffic is generated in a uniform manner in the network. The traffic load in a region must be proportional to the region's area. We define the relative distance σ between two cells as the ratio [41]

$$\sigma = \frac{D}{R} \tag{4.75}$$

where R is the radius of the cell and D is the physical distance between two neighboring cell centers. Owing to practical coverage definition, σ takes its values in the interval $(1, 2)$. We set the parameter σ to the value 1.4.

Traffic intensity will be provided in Erlangs. Assuming that the calls are leaving the system independently from each other according to an exponential distribution with mean $1/\mu$ sec, i.e the average call duration is $\tau = 1/\mu$ sec, the corresponding traffic in Erlangs will be

$$E = \frac{\lambda\tau}{60} \tag{4.76}$$

where the rate λ is given in calls per minute.

The performance measure over which the CLB-DA, LLR and SA algorithms were compared, was the call blocking probability, P_B. In a cellular system, two types of call blocking are generally identified:

- New call blocking probability, which takes place for a newly arrived call, which cannot be accomodated owing to lack of resources.

- Forced call termination probability, which occurs in the case of an ongoing call, which is forced to terminate due to resource shortage or other reason (e.g low signal level)

Usually, the termination of an ongoing call is more significant and therefore the above two kinds of blocking are distinguished. To facilitate study, we treat those identically and consider that a reported blocking event concerns either of the two situations. In order to maintain a common base in our comparison, we selected the call blocking condition to be the same for all policies. This blocking criterion is also consistent with channel reuse constraints in the system.

The performance of CLB-DA, LLR and SA is depicted for a varying number of resources (channels)in figures 4.5 - 4.9. Specifically, 10, 12, 15, 18 or 20 channels are supposed to constitute the system capacity. In all cases we observed that SA (Random Routing) performance provides an upper bound to the performances of CLB-DA and LLR.

Regardless of system capacity, the blocking probability for the SA case grows almost linearly in the semi-logarithmic scale where the graphs are plotted. Therefore, it can be deduced that the blocking probablity P_B grows with the traffic E (in Erlangs) in the following fashion:

$$P_{B,SA} \sim 10^{\rho E} \tag{4.77}$$

where the parameter ρ can be easily calculated from the graphs as the corrsponding slope.

CLB-DA algorithm demonstrates a lower blocking probability than the classical LLR policy in general. Although the cornerstone principle of both algorithms is Load Balancing, consideration of Clique Load Balancing and implementation of sequential assignments leads to a better performance than LLR. For relatively light traffic load, the difference in blocking probability in the two policies is more evident. As traffic increases, both policies converge to the same performance to some extent. We observe however that as system capacity increases, the two policies tend to be more distinguishable.

The performance of the CLB-DA algorithm ,which is of more practical interest here, is a function of system capacity. Generally, for higher capacity, call blocking probability is reduced. CLB-DA can be visualized as very efficient in the case of relatively low offered traffic. Given the system capacity, there exists some finite number E^*, such that if the traffic load E satisfies $E \leq E^*$, then the corresponding blocking probability $P_B(E)$ is negligible. For example, for $N_{CH} = 20$ channels, we have that $P_B(E) < 10^{-3}$ for $E < 6$ Erlangs (4 calls per minute). To guarantee of course that the performance is satisfactory for large enough values of traffic load E, extra capacity has to be sacrificed.

Call QoS standards are predetermined according to various specifications. A valid QoS measure is for example that $P_B < 0.03$. For 10 channels, we have to keep traffic up to the level of 4 Erlangs (1.5 calls/min). Notice however that if we double the capacity (20 channels) traffic can be increased up to 10 Erlangs (about 7 calls/min). The idea of keeping the traffic up to a specified level should not be considered a utopia, since the above analyzed linear cellular system may be a structural part of a greater integrated system, so that traffic can be effectively split between parts of the system.

As traffic increases, CLB-DA performance approaches SA performance asymptotically. This asymptotic behavior becomes more evident for relatively small capacity (up to 15 channels, in our experiment).

Those facts are quite consistent with real environments. Owing to the fact that both CLB-DA and LLR provide a means of handling mobiles in the overlap areas, which constitute a small portion of the entire population, their performance is unavoidably restricted. When traffic intensity is moderate CLB-DA seems to outperform LLR. In the case where traffic load is greater than the system can tolerate, no algorithm handling overlap area mobiles seems to really contribute to any improvement.

4.11.1 Extensions

An approach to the joint problem of base station and channel allocation based to the Load Balancing principle has been considered. The algorithm performs sequential assignments and is shown to converge to a minimum cost solution.

In the above analysis we did not consider the possibility of call handover between cells. While in the overlap area, the mobile can be served by either of the two neighboring base stations. It may be convenient for the call to transition from one base station to the other so that call admission to the base station can be implemented without additional number of channels. This extension should decrease the blocking probability further but will unavoidably introduce additional signalling overhead, owing to handover requests. An obvious extension to the above study would be the consideration of the base station handover and its effect on performance.

Moreover, the combination of other utilized DCA techniques and their integration in the Load Balancing platform and their performance fall within our research scope.

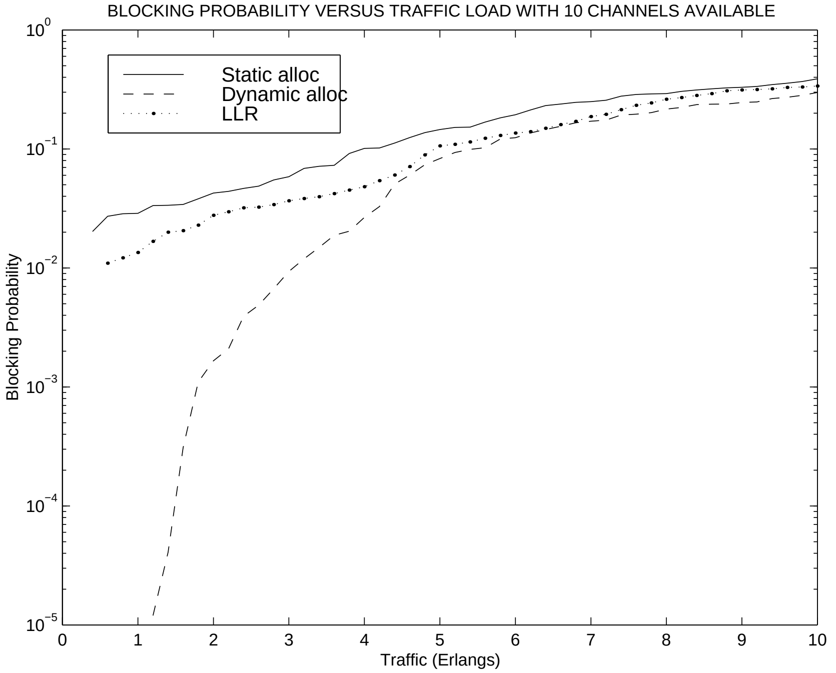

Figure 4.5: Blocking probability vs. offered traffic load for Static and Dynamic allocation and 10 available channels.

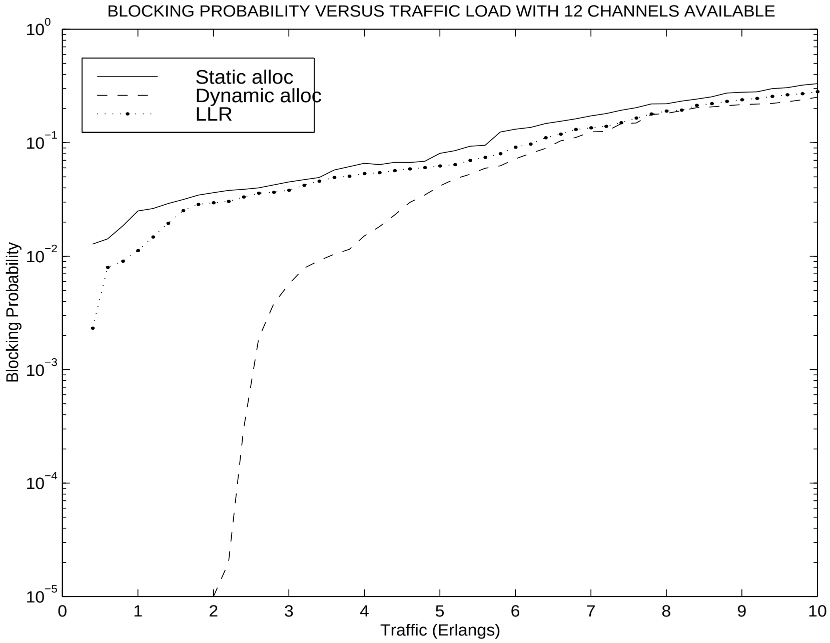

Figure 4.6: Blocking probability vs. offered traffic load for Static and Dynamic allocation and 12 available channels.

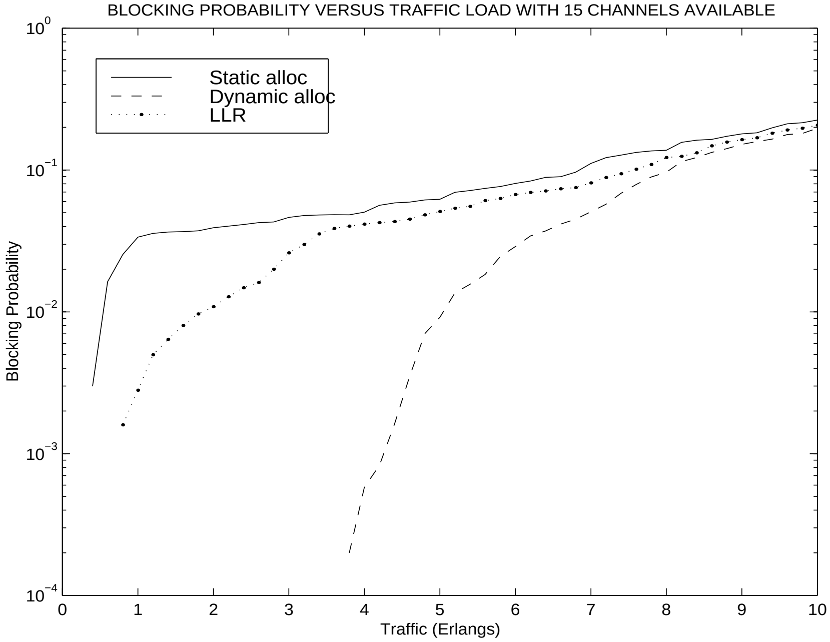

Figure 4.7: Blocking probability vs. offered traffic load for Static and Dynamic allocation and 15 available channels.

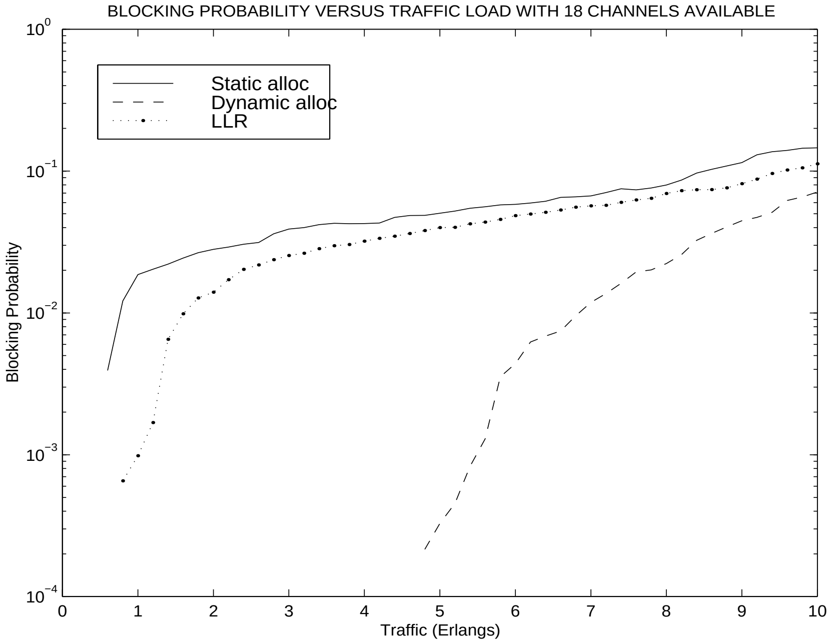

Figure 4.8: Blocking probability vs. offered traffic load for Static and Dynamic allocation and 18 available channels.

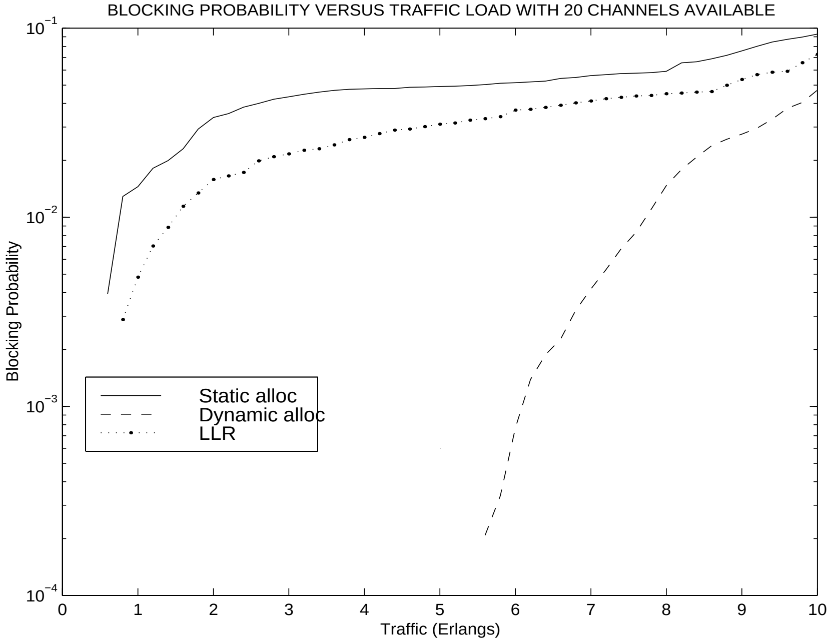

Figure 4.9: Blocking probability vs. offered traffic load for Static and Dynamic allocation and 20 available channels.

Table 4.1: Demonstration of SCLB Algorithm

Iter	$\underline{Q}$ before	$\underline{Q}$ after	$\underline{C}$ after
0	27 31 31 28 26	28 34 37 37 28	11 17 17 20 17 11
1,i=1	28 31 37 37 28	28 31 37 37 28	14 14 17 20 17 11
1,i=2	27 31 32 37 28	33 31 32 37 28	14 19 12 20 17 11
1,i=3	33 31 32 32 28	33 34 32 34 28	14 19 15 17 17 11
1,i=4	33 34 31 34 26	33 34 31 34 29	14 19 15 16 18 11
1,i=5	33 34 31 30 29	33 34 31 30 29	14 19 15 16 14 15
2,i=1	33 34 31 30 29	33 34 31 30 29	14 19 15 16 14 15
2,i=2	27 34 31 30 29	32 34 32 30 29	14 18 16 16 14 15
2,i=3	32 31 32 28 29	32 32 32 32 29	14 18 14 18 14 15
2,i=4	32 32 32 32 26	32 32 32 32 29	14 18 14 18 14 15
2,i=5	32 32 32 32 29	32 32 32 32 29	14 18 14 18 14 15
3,i=1	32 32 32 32 29	32 32 32 32 29	14 18 14 18 14 15
3,i=2	27 32 31 32 29	32 32 32 32 29	14 18 14 18 14 15
3,i=3	32 31 32 28 29	32 32 32 32 29	14 18 14 18 14 15
3,i=4	32 32 32 32 26	32 32 32 32 29	14 18 14 18 14 15
3,i=5	32 32 32 32 29	32 32 32 32 29	14 18 14 18 14 15

Table 4.2: Demonstration of CLBIWF Algorithm

i	cl.	$\underline{Q}$	$\underline{\ell}$	$\underline{C}$	$\underline{\hat{C}}$
1	-	33 40 40 38 29	6 9 9 10 3	10 13 12 14 11 10	4 6 5 3 5
2	2	33 39 40 38 29	6 8 9 9 3	10 13 12 15 11 10	4 6 4 3 5
3	3	33 39 39 38 29	5 8 8 9 3	10 14 12 15 11 10	4 5 4 3 5
4	2	33 38 39 38 29	5 7 8 9 3	11 14 12 15 11 10	3 5 4 3 5
5	3	33 38 38 38 29	4 7 7 9 3	11 15 12 15 11 10	3 4 4 3 5
6	2	33 37 38 38 29	4 6 7 8 3	11 15 12 16 11 10	3 4 3 3 5
7	3	33 37 37 38 29	3 6 6 8 3	11 16 12 16 11 10	3 3 3 3 5
8	4	33 37 37 37 29	3 6 6 7 3	11 16 12 16 11 11	3 3 3 3 4
9	2	33 36 37 37 29	3 5 6 7 3	12 16 12 16 11 11	2 3 3 3 4
10	3	33 36 36 37 29	2 5 5 7 3	12 17 12 16 11 11	2 2 3 3 4
11	4	33 36 36 36 29	2 5 5 6 3	12 17 12 16 11 12	2 2 3 3 3
12	2	33 35 36 36 29	2 4 5 5 3	12 17 12 17 11 12	2 2 2 3 3
13	3	33 35 35 36 29	2 4 4 5 2	12 17 12 17 12 12	2 2 2 2 3
14	4	33 35 35 35 29	2 4 4 4 2	12 17 12 17 12 13	2 2 2 2 2
15	2	33 34 35 35 29	2 3 4 4 2	13 17 12 17 12 13	1 2 2 2 2
16	3	33 34 34 35 29	1 3 3 4 2	13 18 12 17 12 13	1 1 2 2 2
17	4	33 34 34 34 29	1 2 3 3 2	13 18 13 17 12 13	1 1 1 2 2
18	2	33 33 34 34 29	1 1 3 3 2	14 18 13 17 12 13	0 1 1 2 2
19	3	33 33 33 34 29	1 1 2 3 1	14 18 13 17 13 13	0 1 1 1 2
20	4	33 33 33 33 29	1 1 2 2 1	14 18 13 17 13 14	0 1 1 1 1
21	1	32 33 33 33 29	0 1 1 2 1	14 18 14 17 13 14	0 0 1 1 1
22	2	32 32 33 33 29	0 0 1 1 1	14 18 14 18 13 14	0 0 0 1 1
23	3	32 32 32 33 29	0 0 0 1 0	14 18 14 18 14 14	0 0 0 0 1
24	4	32 32 32 32 29	0 0 0 0 0	14 18 14 18 14 15	0 0 0 0 0

Bibliography

[1] http://www.ee.surrey.ac.uk/Personal/L.Wood/constellations/tables/overview.html, *Big LEO overview*

[2] http://www.ee.surrey.ac.uk/Personal/L.Wood/constellations/tables/tables.html, *Big LEO tables*

[3] http://www.teledesic.com *Teledesic official site*

[4] D. Roddy : *Satellite Communications*, Mc Graw-Hill, 1996.

[5] F. Ananasso and F. Delli Priscoli : *"The role of satellites in Personal Communication Services"*, IEEE Journal on Selected areas in Communications, vol. 13, no. 2, pp. 180-196, 1995.

[6] J.V. Evans : *"Satellite systems for Personal Communications"*, IEEE Antennas and Propagation Magazine, vol. 39, no. 3, pp. 7-20, 1997.

[7] M. Asawa and W.E. Stark : *"Optimal scheduling of handoffs in cellular networks"*, IEEE/ACM Transactions on Networking, vol. 4, no. 3, pp. 428-441, 1996.

[8] A. Böttcher and M. Werner : *"Strategies for Handover control in Low Earth Orbit Satellite systems"*, IEEE 44th Vehicular Technology Conference, 1994.

[9] F. Delli Priscoli : *"Functional areas of advanced mobile satellite systems"*, IEEE 47th Vehicular Technology Conference, 1997.

[10] W. Zhao, R. Tafazolli and B.G. Evans : *"Combined handover algorithm for dynamic satellite constellations"*, Electronics Letters, vol. 32, no. 7, pp. 622-624, 1996.

[11] J. Restrepo and G. Maral : *"Coverage concepts for satellite constellations providing communications services to fixed and mobile users"*, Space Communications, vol. 13, no.2, pp. 145-157, 1995.

[12] F. Ananasso and F. Delli Priscoli : *"Technology and Networking issues in 3^{rd} generation satellite personal communication networks"*, Third Annual International Conference on Universal Personal Communications, 1994.

[13] E. Del Re, R. Fantacci and G. Giambene : *"Performance comparison of different Dynamic Channel Allocation techniques for mobile satellite systems"*, European Transactions on Telecommunications, vol. 8, no. 6, pp. 609-621, 1997.

[14] S. Tekinay and B. Jabbari : *"Handover and Channel assignment in mobile cellular networks*, IEEE Communications Magazine, vol.29, no. 11, pp. 42-46, 1991.

[15] G. Ruiz, T. Doumi and J. Gardiner : *"Teletraffic analysis and simulation for nongeostationary mobile satellite systems'*, IEEE Transactions on Vehicular Technology, vol. 47, no. 1, pp. 311-320, 1998.

[16] E. Del Re, R. Fantacci and G. Giambene : *"Efficient Dynamic Channel Allocation techniques with Handover Queuing for mobile satellite networks*,

IEEE Journal on Selected areas in Communications, vol. 13, no. 2, pp. 397-405, 1995.

[17] T. E. Wisloff : *Dual Satellite path diversity and practical channel management for non-geostationary satellite systems*, 1996 5th IEEE International Conference on Universal Personal Communications.

[18] M. Werner, H. Bischl and E. Lutz : *"Handover and satellite diversity in personal satellite communications systems"* EPMCC 1995.

[19] : M. Werner, A. Jahn, E. Lutz and A. Böttcher : *"Analysis of system parameters for LEO/ICO Satellite communication networks"*, IEEE Journal on Selected areas in Communications, vol. 13, no. 2, pp. 371-381, 1995.

[20] Hughes Network Systems : *ICO System Definition*, Draft version 6.0, 1995.

[21] E. Del Re and P. Iannucci : *The GSM procedures in an integrated cellular/satellite system*, IEEE Journal on Selected areas in Communications, vol.13, no.2, pp. 411-430, 1995.

[22] *"The Pan-European cellular system"*, in *The Mobile communications Handbook*, 1995, CRC Press Inc.

[23] M. R. Garey and D.S Johnson : *"Computers and intractability: A guide to the thory of NP-completeness"*, Freeman, New York, 1979.

[24] K. N. Sivarajan, R. J. McEliece and J. W. Ketchum : *"Channel assignment in cellular radio"*, IEEE 39th Vehicular Technology Conference, 1989.

[25] C. Sung and W. Wong : *"A graph theoretic appraoch to the Channel Assignment problem in cellular systems"*, IEEE 45th Vehicular Technology Conference, 1995.

[26] R. Mathar and J. Mattfeldt : *"Channel assignment in Cellular radio networks"*, IEEE Transactions on Vehicular Technology, vol. 42, no. 4, pp. 647-655, 1993.

[27] A. Sen, T. Roxborough and S. Medidi : *"Upper and lower bounds of a class of Channel Assignment problems in Cellular networks"*, IEEE INFOCOM 1998.

[28] M. Sengoku, H. Tamura, S. Shinoda and T. Abe : *"Development in Graph- and/or Network-theoretic research of cellular mobile communication Channel Assignment problems"*, IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, vol. E77-A, no. 7, pp. 1117-1124, 1994.

[29] A. Gamst : *"Some lower bounds for a class of Frequency assignment problems*, IEEE Transactions on Vehicular Technology, vol. 35, no. 1, pp. 8-14, 1986.

[30] V. Prabhu and S. S. Rappaport : *"Approximate analysis for Dynamic Channel Assignment in large systems with cellular structure"*, IEEE Transactions on Communications, vol. 22, no. 10, pp. 1715-1720, 1974.

[31] D. E. Everitt and N. W. Macfadyen : *"Analysis of multi-cellular mobile radio telephone systems with loss"*, British Telecom Journal, vol. 1, no. 2, pp. 34-45, 1983.

[32] F. P. Kelly : *"Stochastic models for computer communication systems"*, Journal of Royal Statistical Socity, vol. 47, no. 3, pp. 379-395, 1985.

[33] P. A. Raymond : *"Performance analysis of cellular networks"*, IEEE Transactions on Communications, vol. 39, no. 12, pp. 1787-1793, 1991.

[34] K. Nakano, M. Sengoku, S. Shinoda and T. Abe : *"Clique Packing and channel assignment in cellular mobile communication systems"*, lectronics and Communications in Japan, Part 1, vol. 79, no. 11, 1996.

[35] R. K. Boel and J. H. van Schuppen : *Distributed Routing for Load Balancing*, Proceedings of the IEEE, vol. 77, no. 1, pp. 210-221, 1989.

[36] M. Alanyali and B. Hajek : *"On simple algorithms for Dynamic Load Balancing"*, IEEE INFOCOM 1995.

[37] B. Hajek : *"Performance of Global Load Balancing by local Adjustment*, IEEE Transactions on Information Theory, vol. 36, no. 6, 1990.

[38] C. Van de Panne : *"Methods for Linear and Quadratic Programming"*, North-Holland Publishing Co., 1975.

[39] W. Zangwill : *"Non-linear Programming : A unified approach*, Prentice-Hall, 1969.

[40] T. M. Cover and J. A. Thomas : *Elements of Information Theory*, Willey, New York, 1991.

[41] T. J. Kahwa and N. Georganas : *A Hybrid Channel Assignment scheme in Large-scale, cellular-structured Mobile Communication systems*, IEEE Transactions on Communications, vol. 26, no. 4, pp. 432-438, 1978.